ABSTRACT

DAVENPORT, CLEMONTINA ALEXANDER. Semiparametric Regression Models for Interacting Covariates. (Under the direction of Arnab Maity.)

Semiparametric regression draws from the advantages of both parametric and nonparametric methods by incorporating previous information into a nonparametric setting. These techniques can speed up computation time and reduce variability, while retaining much of the flexibility of nonparametric models. Interest lies in adequately modeling the effects of interacting predictors on responses of interest, accurately estimating the parameters of the model, and in some cases, testing for the true effects of the predictors and their interaction. In this dissertation several semiparametric techniques for regressing predictors on responses are presented. The unknown and possibly complicated interaction between the predictors are handled in flexible ways, and different estimation and testing procedures are proposed.

Chapter 2 deals with multiple binary responses that are correlated since they are collected from the same individual and may share common predictor effects. In this chapter, we develop a score-based test using a semiparametric modeling framework that jointly models the global effect of the set of predictors. We account for the potentially nonlinear and complicated interaction between predictors using reproducing kernels. Our testing procedure only requires estimation under the null hypothesis and we use multivariate generalized estimating equations (GEEs) to estimate the model components to account for the correlation among the outcomes. We evaluate finite sample performance of our test via simulation study and demonstrated our methods using the CATIE antibody study data.

Chapter 3 utilizes varying coefficient models which allows for complex covariate effects by modeling the regression coefficients as functions of another covariate. For nonparametric varying coefficients, we borrow the idea of parametrically guided estimation to improve asymptotic bias. In this chapter, we develop a parametrically guided estimation procedure for nonparametric varying coefficient models. Local polynomial fitting is used to estimate the unknown model parameters and a method of bandwidth selection via bias-variance tradeoff is proposed. We compare the performance of the guided estimators with their unguided counterparts in both simulation and a real data example.

Chapter 4 extends varying coefficient models to a functional framework. In this chapter, we would like to regress scalar covariates and discrete realizations of an unknown functional covariate on a scalar response while accounting for possible complicated interactions between the predictors of interest. We first propose using a functional varying-coefficient model, but this model can be limited in capturing the interaction effect. Thus we present a more general

and flexible model that can better handle complicated interactions. We propose two global estimation methods and compute the standard errors of our estimators, and present a naive test for the true interaction effect. We compare the linear varying coefficient model and the general model and the estimation methods in simulation and perform limited assessment of the power of our naive test. Finally, we apply our models and estimation schemes to real data.

Semiparametric Regression Models for Interacting Covariates

by
Clemontina Alexander Davenport

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2013

APPROVED BY:

Donald Martin	Ana-Maria Staicu
Jung-Ying Tzeng	Arnab Maity Chair of Advisory Committee

DEDICATION

To Papi, who didn't get to see "his" Ph.D.

BIOGRAPHY

Clemontina Alexander Davenport was born in Tallahassee, FL to Winfred and Shérre Alexander. She graduated from Leon High School in 2002 and attended Saint Augustine's College (now University) in Raleigh, NC where she graduated summa cum laude in 2006 with a B.S. in Mathematics. She attended Hampton University in Hampton VA, earning an M.S. in Applied Mathematics in 2008 after completing her thesis, *Using the Hamming distance for nonparametric message detection*, under the guidance of Drs. Carolyn and Morris Morgan. She returned to Raleigh to attend NC State University where she earned a Master of Statistics degree in 2010.

ACKNOWLEDGEMENTS

First and foremost, I'd like to acknowledge my advisor Arnab Maity for taking a gamble on a small fish in a big pond. If not for his friendly and easy-going attitude, I'm sure the research process could have been a lot more stressful than it was. I'd like to thank my committee members, Jung-Ying Tzeng, Donald Martin, and Ana-Maria Staicu. I want to thank Jackie Hughes-Oliver and Kim Weems for getting me back on track when I crashed and burned (both times) and their never-ending support, Adrian Blue for always making sure the finances were right, Alison McCoy, and all the other faculty, staff, and students who have helped and supported me.

I want to acknowledge my thesis advisors at Hampton, Carolyn Morgan who would hand me tissues at the same time as telling me to suck it up, and Morris Morgan who was always available to answer dumb questions and dismiss my constant anxiety. From undergrad, I want to thank Diane Suber, Gloria Payne, and all my professors who helped me get so far.

I'd like to acknowledge my awesome husband, Eric, who was right there in the trenches with me from the beginning (struggling through ST522 together), and my Mami; words can not express how grateful I am for her love and support these past 30 years. I'm doing my best to make her proud. I also want to acknowledge my Papi who always encouraged me in his own funny way, Aunt Cookie and Uncle Lewis for being my home away from home and sometimes my literal home, Bryan for keeping me motivated by his lifelong goal to be better than me at everything, Recey, our friendship makes me tear up sometimes (like right now, as I'm writing this), the Alexanders, the Bakers, the Hornes, the Schwartzes, the Davenports, and all of my other family and friends who have encouraged and supported me through this journey. *When one of us makes it, we all make it!*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

A typical regression problem involves determining the effects of some predictor variables on response variables of interest. In parametric regression modeling, the predictors are assumed to have a pre-specified effect (e.g. linear) on the responses and these effects are captured in a finite parameter vector. Analysis is geared towards estimating the parameter vector and testing hypotheses about the true effects of the predictors based on the estimate of the parameter vector. These types of models are simple to construct and interpret, and when the assumptions are met, provide excellent results and large statistical power. However, when the assumptions are not true, serious bias can arise in the estimates and any conclusions drawn from the analysis will be suspect.

Another drawback of parametric modeling occurs when the values of some predictors can change depending on the values of others, deemed interaction between the predictors. Parametric approaches can be very limited in modeling interaction effects and the number of corresponding parameters can increase dramatically as covariates and interaction terms are added. If the unknown interaction between the predictors is somewhat complicated, then parametric techniques will fail at capturing the true relationship. Thus more flexible modeling techniques are needed for real world applications that do not fit the basic textbook situations.

Nonparametric regression is a robust alternative to parametric approaches. The relationship between the predictors and the responses is not fixed beforehand and the data are allowed to "speak for themselves". In many nonparametric models, the predictors are regressed on the response by unknown functions that need to be estimated, and these types of models are the focus of this dissertation. These estimators tend to be more variable and slower to converge than their parametric counterpart, however they do not require as strong assumptions and allow for more modeling flexibility. In many cases, one can obtain the advantages of both parametric and nonparametric regression by combining some previous information or assumptions about part

of the data into a nonparametric setting. For instance, if a subset of covariates of interest is historically known to have a certain effect, then these predictors can be modeled parametrically, while the rest take a nonparametric form. These types of semiparametric techniques can speed up computation time and reduce variability.

Different semiparametric regression models for interacting covariates are considered in this dissertation. The main focus is on adequately modeling the interaction between predictors, accurately and efficiently estimating the parameters of the models, and in some cases, testing for the true effects of the predictors. Each chapter and subsequent methodology was motivated by data that were not well suited for parametric techniques. In Chapter 2, the data consist of multiple correlated, dichotomous responses that were influenced by sets of predictors that interacted in an unknown and possibly complicated way. A procedure to test for the effect of the entire set of predictors is proposed, while accounting for the interaction between the predictors in the set, and the correlation between the multiple responses. The testing procedure only requires estimation under the null hypothesis and multivariate generalized estimating equations are used to estimate the model components.

Chapter 3 demonstrates how nonparametric varying coefficient model estimators can be enhanced by incorporating prior knowledge of the data in the form of a parametric start. Varying coefficient models are a flexible extension of linear models that can handle complicated interaction effects, but estimates of the unknown functions can be biased if the curvature in the true functions is too sharp. If some insight into the shape of the unknown function is given, then one can identify a parametric family that captures that shape, and then remove the shape from the original function. Standard nonparametric local polynomial fitting can then be applied to the remaining flat function, after which the shape is added back to obtain the final estimates. The bandwidth parameter from local smoothing must be estimated, and we propose doing this using a bias-variance tradeoff. The guided estimators allow us to choose a higher bandwidth without losing the bias advantage and this results in smaller variation as well. These enhancements are presented in simulations and the methodology is applied to a real data example.

Lastly, in Chapter 4 semiparametric models are presented for situations in which one of the predictors of interest is functional and can interact with a scalar predictor. Interest lies in regressing the scalar covariate of interest and the functional covariate while accounting for the possibly complicated interaction between the two. An extension of the varying coefficient model to the functional framework is proposed, but is limited in interaction fitting. In contrast, a more general and flexible model is presented as an alternative. The two estimation schemes proposed in this chapter allows for the variability of the estimators to be computed and as a result, confidence bands and surfaces can be constructed. The interaction between the functional

and scalar predictor is flexibly modeled by an unknown, bivariate functional parameter, and a simple test for interaction effect is proposed. The different models and estimation procedures are compared via simulation and real data.

Chapter 2

A Powerful Test for SNP Effects on Multivariate Binary Outcomes

2.1 Introduction

Analyzing multiple binary outcomes collectively is a common problem in studies of complex diseases. For example, in cardiac studies, the joint assessment of abnormal blood pressure, reduced pumping ability, and cardiomyopathy together can give a more complete picture of cardiac function than analyzing each of these factors individually (Lipsitz et al., 2009). In the Maternal Life Study (Bauer et al., 2002), multiple binary central and autonomic nervous system signs were measured on infants exposed to cocaine prenatally. Together, these signs are important in determining if the newborn displayed narcotic withdrawal syndrome (Das et al., 2004). Likewise, in genetic studies, often multiple binary phenotypes from each individual are collected and these phenotypes combined are relevant to the disease of interest. We consider the situation where genetic information on a particular set of single nucleotide polymorphisms (SNPs) of interest is also collected for each individual and the primary interest lies in determining whether the marker set has any association with the observed multiple binary outcomes. In this chapter, our primary goal is to develop a global test for association between the marker set of interest and multiple, possibly correlated, binary outcomes. Our motivating data comes from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) antibody study (discussed in detail in Section 2.4), where the presence or absence of antibodies to three neurotrophic herpesviruses were measured in individuals with schizophrenia. In addition, genotype data are available for 492K SNPs (Sullivan et al., 2008). Primary challenges in developing association tests for this kind of data are to incorporate possible correlation between the multiple binary outcomes and to account for complex and possibly nonlinear interactions among the markers in question. We

develop a testing procedure that accounts for both of these issues.

We first discuss incorporating possibly complex interactions between the markers in the testing procedure. It is well known that genetic markers do not act alone, rather they act together in a network and the resulting pathway effect may cause phenotypic changes such as disease occurrence. Thus, it is usually better to study the effects of the whole marker set rather than each individual marker since there are many drawbacks of performing individual marker analysis. The three major issues with individual analysis are: 1) the results are often not reproducible because the selection of significant markers is arbitrary leading to different biological conclusions (Pan et al., 2005), 2) single marker analysis often results in reduced power due to multiple comparisons, and 3) since the markers are analyzed individually, interaction between them cannot be studied (Nam and Kim, 2008). Many marker set analysis methods have been recently developed (see e.g., Nam and Kim, 2008; Kwee et al., 2008; Wu et al., 2011) to account for such possible interactions. To be specific, nonparametric regression methods based on kernel machine regression (KMR) have become very popular in association testing problems because of their ability to incorporate complex interactions (Liu et al., 2007; Kwee et al., 2008; Liu et al., 2008). In the KMR framework, the joint effect of the marker set is modeled using a positive definite kernel that essentially computes pairwise similarity measures between individuals and attempts to associate them to the outcome variable at hand. Because no parametric assumptions are made on the joint marker set effect, it is advantageous over the typical approach of fitting a parametric generalized linear model (GLM) where the group effect is modeled as a linear combination of the individual SNPs. Parametric methods are easier to fit but are typically restrictive because most of the time there is complex interaction between the SNPs that a linear model cannot account for. Even for a small number of genes, the number of interaction parameters can be large and cause dimensionality issues. Also, violating the linearity assumption may result in loss of power. Thus, kernel based methods offer an attractive and powerful alternative to develop association tests. Support vector machines (Vapnik, 1998) and their least squares extension (Suykens et al., 2002) are common examples of kernel methods. We direct readers to Hofmann et al. (2008) for an overview of the developments of this method in machine learning.

The elegance of kernel methods lies in the fact that using a positive definite kernel allows hypothesis testing in a GLM framework while still accounting for nonlinear dependence and a large number of SNPs. However, to the best of our knowledge, all of these methods are developed for the case when one has a single response variable. For example, Liu et al. (2007), Kwee et al. (2008), and Liu et al. (2008) use KMR to model the gene pathway effects on a single, phenotypic, response while accounting for other covariates. These papers do not address the common case when multiple phenotypes for each individual are observed, such as in the

CATIE data. Recently, Maity et al. (2012) used KMR for multiple phenotypes and develop a score-based test statistic, but focus solely on continuous responses. They use likelihood-based techniques that cannot be directly applied to our case, since the joint likelihood of correlated binary outcomes is difficult to specify. In this chapter, we are interested in developing a testing procedure in presence of multiple binary outcomes while accounting for complex interactions among the markers.

One of the main challenges of analyzing multivariate binary responses together in a GLM framework is that the correlation structure is usually unknown and the joint distribution of the outcomes is difficult to construct. Thus a direct development of score-based testing procedures as in KMR with continuous outcomes is not straightforward. We will develop our testing procedure by synergistic use of nonparametric KMR and the generalized estimating equation (GEE) frameworks. When studying multiple phenotypes, correlation can arise for two reasons: 1) the responses are collected on the same individual, and 2) the same set of markers can have effect on more than one response variable. A typical (but naive) approach to analyzing such a data set is to consider one outcome at a time, use an association testing procedure of choice (see e.g., Liu et al., 2008), then perform a Bonferroni correction on the resulting p-values. However this procedure completely ignores any correlation structure between the responses, and thus may suffer from loss of power. Addressing both sources of correlation could result in increased power to detect genetic effects when there is moderate to high correlation among the outcomes, which will be demonstrated in simulations. We address this problem by using generalized estimating equations (GEEs) to estimate the parameters of the model with a user-specified correlation structure. Even if the correlation structure is misspecified, which is likely in practical situations, GEE estimators are still robust and consistent, and the joint distributions do not need to be specified (Liang and Zeger, 1986). We develop a score-type test that is advantageous because estimation only needs to be done under the null model. We show in simulation that our proposed method is comparable in performance to naively testing the outcomes individually when there is low correlation among the outcomes, but is more powerful when the correlation is moderate or high.

The rest of this Chapter is organized as follows. Section 2.2 provides a review of kernel machine regression for the univariate response model and then describes our proposed testing procedure. In Section 2.3, we evaluate the performance of our proposed procedure via simulation study. We find that as both sources of correlation increase, our method is much more powerful than naively evaluating the outcomes separately, while maintaining proper Type I error rate. We then apply our test statistic to the CATIE antibody data in Section 2.4. Finally we provide some discussion and concluding remarks in Section 2.5.

2.2 Methodology

We first review the KMR testing methodology when there is only one binary outcome and its connection to penalized quasi-likelihood mixed modeling framework. Then we will develop our testing procedure when there are multiple binary outcomes observed for each sampling unit.

2.2.1 Kernel Machine Association Testing for Univariate Binary Outcome

Assume that for each of n subjects we have a univariate binary outcome, demographic covariates such as age and sex, and genetic information such as the number of minor alleles at each SNP on a chromosome. Let Y_i be the binary response for ith individual, $i = 1, ..., n$, with corresponding $q \times 1$ covariate vector $\mathbf{X}_i$ and $p \times 1$ vector $\mathbf{Z}_i$ of SNP information. Define $\mathbf{Y} = (Y_1, ..., Y_n)^T$ and $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)^T$. The primary goal is to study the effects of $\mathbf{Z}$ on $\mathbf{Y}$ after adjusting for $\mathbf{X}$. Because genes may interact in an unknown and complicated way, Liu et al. (2008) proposed the model

$$\text{logit}(p_i) = \mathbf{X}_i^T \boldsymbol{\beta} + h(\mathbf{Z}_i), \tag{2.1}$$

where $p_i = \text{P}(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i)$, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $h(\cdot)$ is an unknown scalar function. By modeling the covariates parametrically and the gene effect nonparametrically, this model is flexible in the genetic effects and allows for nonlinearity and complicated interaction of SNPs in the same pathway.

In the standard KMR for binary outcomes, we assume that h lies in a function space $\mathcal{H}_k$. This space is uniquely specified by a symmetric, positive definite kernel function, $K(\cdot, \cdot)$, and is spanned by a set of orthogonal basis functions of the form $\Phi(\mathbf{z}) = \{\phi_j(\mathbf{z})\}_{j=1}^J$ (Liu et al., 2007). Typically $\mathcal{H}_k$ is infinite dimensional and the actual basis functions are unknown, thus it is difficult to work with Φ directly. Alternatively, we set the kernel function equal to the inner product defined on $\mathcal{H}_k$, that is, $K(\mathbf{z}_i, \mathbf{z}_j) = \langle \Phi(\mathbf{z}_i), \Phi(\mathbf{z}_j) \rangle_{\mathcal{H}_k}$. This way, instead of representing all $h(\mathbf{z}) \in \mathcal{H}_k$ using the primal representation $h(\mathbf{z}) = \sum_{j=1}^J \omega_j \phi_j(\mathbf{z}) = \Phi(\mathbf{z})^T \boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is a vector of coefficients, we can use the dual representation $h(\mathbf{z}) = \sum_{l=1}^n \alpha_l K(\mathbf{z}_l, \mathbf{z}; \rho)$ for some constants α_l. This representation is more convenient to work with because the explicit basis functions do not need to be specified (Burges, 1998; Liu et al., 2007).

The choice of $K(\cdot, \cdot)$ determines which function space is used to approximate the unknown h. For instance, a dth order polynomial kernel $K(\mathbf{z}_1, \mathbf{z}_2) = (\mathbf{z}_1^T \mathbf{z}_2 + 1)^d$ generates the function space spanned by the basis functions $\phi_j(\mathbf{z}) = \{z_k^a z_{k'}^b : a, b \in \mathbb{N}, a + b \leq d\}$ $(k, k' = 1, ..., p)$ and thus, $d = 1$ and $d = 2$ means we assume $h(\cdot)$ is linear and quadratic in the z's, respectively. The function space generated by the Gaussian kernel $K(\mathbf{z}_1, \mathbf{z}_2) = \exp\{-||\mathbf{z}_1 - \mathbf{z}_2||^2/\rho\}$ is spanned by radial basis functions (Buhmann, 2003), where ρ is a tuning parameter. There are many

kernel options to choose from depending on the problem at hand (Hofmann et al., 2008) and in our framework, we can view $K(\cdot,\cdot)$ as a measure of how similar two individuals are in terms of their genotype information. In marker set analysis, a popular kernel for SNP data that allows for SNP interaction (epistasis) is the IBS kernel defined as

$$K(z_i, z_j) = \sum_{s=1}^{p} \mathrm{IBS}(z_{i,s}, z_{j,s}) \Big/ 2p, \tag{2.2}$$

where $\mathrm{IBS}(z_{i,s}, z_{j,s})$ is the number of alleles shared identical by state (IBS) by individual i and j at SNP s, $s = 1, ..., p$ (Kwee et al., 2008). Wu et al. (2011) explain that because the number of alleles shared IBS does not depend on different genotype encodings, using the IBS kernel removes the assumption of additivity found in many genetic models. They also suggest using this kernel when the amount of interaction is modest.

The parameters $\boldsymbol{\beta}$ and h in (2.1) can be estimated by maximizing the penalized log-likelihood using a Fisher scoring or a Newton-Raphson algorithm. Liu et al. (2008) show that, by treating $\boldsymbol{\beta}$ as a vector of fixed effects and $\mathbf{h} = (h_1, ..., h_n)^T$ as a vector of random effects, the logistic KM estimator is the same as the penalized quasi-likelihood estimator from a logistic mixed model $\mathrm{logit}(p_i) = \mathbf{X}_i^T\boldsymbol{\beta} + h_i$, where $\mathbf{h} \sim N(\mathbf{0}, \tau\mathbf{K})$, $\tau = 1/\lambda$, λ is the penalty parameter from the penalized likelihood, and $\mathbf{K}$ is a square matrix whose (i,j)th element is (2.2). The normal equations given in (5) of Liu et al. (2008) coincides with iteratively fitting a working linear mixed model

$$\tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\varepsilon}$$

until convergence where $\boldsymbol{\beta}$ and $\mathbf{h}$ are estimated using BLUE and BLUP respectively and $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$ where $\mathbf{D} = \mathrm{diag}\{p_i(1-p_i)\}$. The regularization parameter τ can be estimated by treating it as a variance component and maximizing the REML criterion

$$\ell \approx -\frac{1}{2}\log|\mathbf{V}_u| - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}_u^{-1}\mathbf{X}| - \frac{1}{2}(\tilde{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{V}_u^{-1}(\tilde{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{2.3}$$

where $\mathbf{V}_u = \mathbf{D}^{-1} + \tau\mathbf{K}$ and $\tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \mathbf{D}^{-1}(\mathbf{Y} - \mathbf{p})$. We refer to Liu et al. (2008) for full details.

Testing the overall genetic effect $H_0\colon h(\mathbf{z}) = 0$ for univariate responses is equivalent to testing $H_0\colon \tau = 0$. Liu et al. (2008) propose the following score test statistic based on the derivative of (2.3) with respect to τ

$$S = \frac{Q(\hat{\boldsymbol{\beta}}_0) - p_Q}{\sigma_Q}, \tag{2.4}$$

where $Q(\hat{\boldsymbol{\beta}}_0) = (\tilde{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_0)^T \mathbf{DKD}(\tilde{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_0) = (\mathbf{Y} - \hat{\mathbf{p}}_0)^T \mathbf{K}(\mathbf{y} - \hat{\mathbf{p}}_0)$, $\text{logit}(\hat{\mathbf{p}}_0) = \mathbf{X}\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}_0$ is the MLE of $\boldsymbol{\beta}$ under the null logistic model, $p_Q = \text{tr}\{\mathbf{P}_0\mathbf{K}\}$, $\sigma_Q = 2\text{tr}\{\mathbf{P}_0\mathbf{K}\mathbf{P}_0\mathbf{K}\}$, and $\mathbf{P}_0 = \mathbf{D}_0 - \mathbf{D}_0\mathbf{X}(\mathbf{X}^T\mathbf{D}_0\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}_0$ where $\mathbf{D}_0 = \text{diag}\{\hat{p}_{i0}(1 - \hat{p}_{i0})\}$.

Two advantages of this test statistic are 1) the basis functions for $h(\cdot)$ do not need to be specified, which is often a difficult task, and 2) their test has more power than tests based on a parametric assumption. Liu et al. (2008) compared their test to a global test based on the linearity assumption and showed in simulations that their test was as powerful as the global test and suggest using their test as a universal test for both linear and nonlinear pathway effect.

2.2.2 Testing with Multivariate Binary Outcomes

In this section, we are interested in situations where multiple binary outcomes are measured on each subject. Let Y_{ij} be the jth observation for the ith individual, $i = 1, ..., n$ and $j = 1, ..., t$ so that the response for a given individual i is $(Y_{i1}, ..., Y_{it})$. Define the covariates and SNP information corresponding to Y_{ij} as $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively. Multiple outcomes are observed on each subject, but the covariates (e.g., age, sex, etc.) and SNPs do not change over j. Denote $\mathbf{Y}_j = (Y_{1j}, Y_{2j}, ..., Y_{nj})^T$ as the $n \times 1$ vector of all of the jth observations. Because we are grouping the outcomes by observation and not by individual, all of the elements in $\mathbf{Y}_j$ are independent of each other but the vectors themselves are not. We combine the covariates in a similar way by defining $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)^T$ as the $n \times q$ matrix of covariates for the jth observations. We assume the random variable Y_{ij} has a Bernoulli distribution with mean $p_{ij} = \text{E}(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i)$ and variance $\text{Var}(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i) = p_{ij}(1 - p_{ij})$. Marginally, for each $j = 1, ..., t$, we can write the model

$$\text{logit}(p_{ij}) = \mathbf{X}_i^T \boldsymbol{\beta}_j + h_j(\mathbf{Z}_i). \tag{2.5}$$

Here $\boldsymbol{\beta}_j$ and h_j indicate that the covariates and SNPs do not necessarily have the same effect on the different responses of each individual. We would like to determine the global effect of the SNPs by testing H_0: $\mathbf{h}(\cdot) = \{h_1(\cdot), ..., h_t(\cdot)\}^T = \mathbf{0}$.

Test statistic

A naive approach to our testing problem would be to fit the marginal model (2.5) to each $\mathbf{Y}_j$ separately and then test for the significant genetic effect using (2.4) and adjust for multiple testing, but two issues arise. The first is that the multiple outcomes on a particular individual are correlated and this correlation is not being accounted for. The second is that performing multiple comparison adjustments can result in reduced power. Alternatively, one could use a combined test statistic

$$Q_{\text{naive}}(\hat{\boldsymbol{\beta}}) = (\mathbf{Y} - \hat{\mathbf{p}})^T \mathbf{K}^* (\mathbf{Y} - \hat{\mathbf{p}}), \tag{2.6}$$

where $\mathbf{Y} = (\mathbf{Y}_1^T, ..., \mathbf{Y}_t^T)^T$, $\hat{\mathbf{p}} = (\hat{\mathbf{p}}_1^T, ..., \hat{\mathbf{p}}_t^T)^T$ with $\text{logit}(\hat{\mathbf{p}}_j) = \mathbf{X}\hat{\boldsymbol{\beta}}_j$, and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, ..., \hat{\boldsymbol{\beta}}_t^T)^T$ is the vector of estimated regression parameters, and $\mathbf{K}^*$ is an $nt \times nt$ block diagonal matrix with blocks $\mathbf{K}$. The test statistic in (2.6) combines all the residuals $\mathbf{Y}_j - \mathbf{p}_j$ and attempts to build a single test to bypass the issue of multiple testing correction. However, Q_{naive} also ignores any correlation structure present among the multiple outcomes and while this strategy works well when the outcomes are truly independent, testing using Q_{naive} results in inflated Type I errors when there is correlation among the outcomes. In a simulation study with $n = 200$ subjects, when the $t = 3$ outcomes per subject were generated to have a compound symmetric correlation structure with correlation 0.5, the Type I error for testing with Q_{naive} was 0.060, and when the correlation was 0.7, the Type I error was 0.079. The more correlated the multiple outcomes are, the more inflated the Type I error becomes and thus Q_{naive} is not a reliable choice for a test statistic.

We develop a method for jointly analyzing the multiple binary responses while taking into account the correlation between them but still keeping the flexible modeling of the SNPs. To do this, we utilize the connection of KMR to the mixed model framework but extend it to the case of multiple responses. First, note that fitting the KMR model for each outcome individually is equivalent to iteratively fitting the working model

$$\begin{bmatrix} \tilde{\mathbf{Y}}_1 \\ \vdots \\ \tilde{\mathbf{Y}}_t \end{bmatrix} = \begin{bmatrix} \mathbf{X}\boldsymbol{\beta}_1 \\ \vdots \\ \mathbf{X}\boldsymbol{\beta}_t \end{bmatrix} + \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_t \end{bmatrix}, \tag{2.7}$$

where $\tilde{\mathbf{Y}}_j = \mathbf{X}\boldsymbol{\beta}_j + \mathbf{h}_j + \mathbf{D}_j^{-1}(\mathbf{Y}_j - \mathbf{p}_j)$, $\mathbf{h}_j = (h_{1j}, ..., h_{nj})^T$, $\boldsymbol{\varepsilon}_j = (\varepsilon_{1j}, ..., \varepsilon_{nj})^T$ for $j = 1, ..., t$, $(\mathbf{h}_1^T, ..., \mathbf{h}_t^T)^T \sim \text{MVN}(\mathbf{0}, \mathbf{V}_h)$ where $\mathbf{V}_h$ is a block diagonal matrix with $\tau_j\mathbf{K}$ as the blocks, and the $\boldsymbol{\varepsilon}_j$'s are independent with variance-covariance matrices $\mathbf{D}_j$, respectively. This, however does not take into account any correlation between the multiple outcomes. In practice, the true correlation structure among the outcomes is unknown, so we posit a working correlation structure for $(\varepsilon_{i1}, ..., \varepsilon_{it})$ in the form of $\mathbf{G}(\boldsymbol{\theta})$, where $\mathbf{G}$ is a matrix that is known up to a parameter $\boldsymbol{\theta}$. By constructing the vector of errors as $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, ..., \boldsymbol{\varepsilon}_t^T)^T$, the working correlation of all the errors for all the responses $\boldsymbol{\varepsilon}$ has the form $\mathbf{S} = \mathbf{G}(\boldsymbol{\theta}) \otimes \mathbf{I}_n$, where $\otimes$ denotes the Kronecker product and $\mathbf{I}_n$ is an $n \times n$ identity matrix. Thus, if we were to use a working unstructured correlation, then $\mathbf{G}(\boldsymbol{\theta})$ has 1's along the diagonal and $\mathbf{G}(\boldsymbol{\theta})_{jk} = \theta_{jk}$ when $j \neq k$.

Define $\mathbf{D}$ as the block diagonal matrix with blocks $\mathbf{D}_1, ..., \mathbf{D}_t$. Then the variance-covariance matrix of $\boldsymbol{\varepsilon}$ is $\mathbf{D}^{-1}\mathbf{D}^{1/2}\mathbf{S}\mathbf{D}^{1/2}\mathbf{D}^{-1} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ and the modified working model will have the same form as (2.7) but with $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2})$. The parameters $\boldsymbol{\beta}_j$ and $\mathbf{h}_j$ can now be estimated using BLUE and BLUP respectively, and the variance components τ_j can be

estimated by maximizing the restricted quasi-likelihood criterion

$$\ell \approx -\frac{1}{2}\log|\mathbf{V}_m| - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}_m^{-1}\mathbf{X}| - \frac{1}{2}\begin{pmatrix}\tilde{\mathbf{Y}}_1 - \mathbf{X}\hat{\boldsymbol{\beta}}_1\\ \vdots \\ \tilde{\mathbf{Y}}_t - \mathbf{X}\hat{\boldsymbol{\beta}}_t\end{pmatrix}^T \mathbf{V}_m^{-1}\begin{pmatrix}\tilde{\mathbf{Y}}_1 - \mathbf{X}\hat{\boldsymbol{\beta}}_1\\ \vdots \\ \tilde{\mathbf{Y}}_t - \mathbf{X}\hat{\boldsymbol{\beta}}_t\end{pmatrix}, \tag{2.8}$$

where $\mathbf{V}_m = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2} + \mathbf{V}_h$.

Our main goal is to test for genetic pathway effects on the multiple binary outcomes $H_0\colon \mathbf{h}(\cdot) = \mathbf{0}$ which is now equivalent to testing $H_0\colon \tau_1 = \cdots = \tau_t = 0$. To do this, we propose a score-type test statistic based on the derivative of the quasi-likelihood like that in (2.8). Taking the derivative of the criterion in (2.8) with respect to τ_j for $j = 1, ..., t$ and then setting $\tau_j = 0$, the score function for τ_j is $S_j = Q_j(\boldsymbol{\beta}, \boldsymbol{\theta}) - p_{jQ}$, where

$$Q_j(\boldsymbol{\beta}, \boldsymbol{\theta}) = (\mathbf{Y} - \mathbf{p})^T \left(\mathbf{S}\mathbf{D}^{1/2}\right)^{-1} \mathbf{D}^{1/2}\mathbf{K}_j^*\mathbf{D}^{1/2} \left(\mathbf{D}^{1/2}\mathbf{S}\right)^{-1} (\mathbf{Y} - \mathbf{p}),$$

where $\mathbf{K}_j^*$ is a block diagonal matrix with the jth block equal to $\mathbf{K}$ and all other blocks $\mathbf{0}$, and $p_{jQ} = \mathrm{tr}\{\mathbf{P}\mathbf{K}_j^*\}$. Because the τ_j's are considered as variance components and thus are nonnegative, testing $H_0\colon \tau_1 = \cdots = \tau_t = 0$ is equivalent to testing $H_0\colon \tau_1 + \cdots + \tau_t = 0$. We can now adopt a similar technique to that of Maity et al. (2012) to obtain our final test statistic as

$$Q(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = (\mathbf{Y} - \hat{\mathbf{p}})^T \left(\hat{\mathbf{S}}\hat{\mathbf{D}}^{1/2}\right)^{-1} \hat{\mathbf{D}}^{1/2}\mathbf{K}^*\hat{\mathbf{D}}^{1/2} \left(\hat{\mathbf{D}}^{1/2}\hat{\mathbf{S}}\right)^{-1} (\mathbf{Y} - \hat{\mathbf{p}}), \tag{2.9}$$

where the hats on $\hat{\mathbf{p}}$, $\hat{\mathbf{S}}$, and $\hat{\mathbf{D}}$ indicate the evaluation of these matrices at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$. Recall that $\mathbf{S} = \mathbf{G}(\boldsymbol{\theta}) \otimes \mathbf{I}_n$ and hence if no correlation is present, then $\mathbf{S}$ becomes an $nt \times nt$ identity matrix and our statistic reduces to Q_{naive} in (2.6).

We chose a logistic model for the conditional probability but a probit or some other model could be substituted. Our test statistic still has the advantages of that in Liu et al. (2008) but also accounts for correlation in the multiple responses.

Estimation under null hypothesis

In order to evaluate Q, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ must be estimated under H_0. To do this, we use the generalized estimating equation (GEE) framework (Liang and Zeger, 1986) which is a multivariate extension of the quasi-likelihood approach discussed in detail by Agresti (2002). GEEs are advantageous over likelihood based methods because the joint likelihood for the multiple binary outcomes is difficult to construct and in the GEE framework, this joint likelihood does not need to be specified. Also, in practice, the true structure of $\mathbf{G}(\boldsymbol{\theta})$ is unknown and a working or posited

structure is used instead. Contrary to likelihood based methods, if the correlation structure is misspecified, the GEE estimators will still be unbiased and robust. GEE estimators are asymptotically normal and consistent, where consistency depends only on a correctly specified model and not the correlation structure $\mathbf{G}(\boldsymbol{\theta})$ (Liang and Zeger, 1986).

We posit the GEEs under H_0

$$\tilde{\mathbf{X}}^T \mathbf{D} \mathbf{V}_{m0}^{-1} (\mathbf{Y} - \mathbf{p}) = \mathbf{0}, \tag{2.10}$$

where $\tilde{\mathbf{X}}$ is an $nt \times nt$ block diagonal matrix with elements $\mathbf{X}$ and $\mathbf{V}_{m0} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$. If there is no genetic pathway effect (H_0: $\mathbf{h}(\cdot) = \mathbf{0}$ is true), then $\mathbf{V}_{m0}$ is the working variance-covariance matrix of $\mathbf{Y}$. To solve (2.10), Liang and Zeger (1986) suggest using a modified Fisher scoring algorithm to find $\boldsymbol{\beta}$ and a method of moments estimation for $\boldsymbol{\theta}$. The updating equation is

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + (\tilde{\mathbf{X}}^T \hat{\mathbf{D}} \hat{\mathbf{V}}_{m0}^{-1} \hat{\mathbf{D}} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \hat{\mathbf{D}} \hat{\mathbf{V}}_{m0}^{-1} (\mathbf{Y} - \hat{\mathbf{p}}).$$

The initial estimates in the first iteration come from fitting a generalized linear model assuming independence.

Null distribution of Q

We use a simulation-based technique to get the p-values of our test statistic in (2.9). For ease of notation, we omit the hats on all the matrices that depend on the estimated parameters. Define $g(\hat{\boldsymbol{\beta}}_j) = e^{\mathbf{X}\hat{\boldsymbol{\beta}}_j}/(1 + e^{\mathbf{X}\hat{\boldsymbol{\beta}}_j})$. The variance of $\mathbf{Y}_j - \hat{\mathbf{p}}_j$ can be computed by using a Taylor series approximation for g and computing the variance of $\mathbf{Y}_j - g(\hat{\boldsymbol{\beta}}_j)$. It can be shown that $\text{var}(\mathbf{Y}_j - \hat{\mathbf{p}}_j) = \mathbf{D}_j - \mathbf{D}_j\mathbf{X}(\mathbf{X}^T\mathbf{D}_j\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}_j \stackrel{\text{def}}{=} \mathbf{P}_j$. If the multiple outcomes were independent, then the variance-covariance matrix of $\mathbf{Y} - \hat{\mathbf{p}}$ would be $\mathbf{P}$, a block diagonal matrix with elements $\mathbf{P}_j$, but since the outcomes are correlated, the variance-covariance matrix of $\mathbf{Y} - \hat{\mathbf{p}}$ is $\mathbf{P}^{1/2}\mathbf{S}\mathbf{P}^{1/2}$ which is no longer block diagonal. By defining $\mathbf{M} = \mathbf{P}^{1/2}\mathbf{S}\mathbf{P}^{1/2}$, (2.9) can be rewritten as $Q = \{\mathbf{M}^{-1/2}(\mathbf{Y} - \hat{\mathbf{p}})\}^T \mathbf{B} \{\mathbf{M}^{1/2}(\mathbf{Y} - \hat{\mathbf{p}})\}$ where $\mathbf{B} = \mathbf{M}^{1/2}(\mathbf{S}\mathbf{D}^{1/2})^{-1}\mathbf{D}^{1/2}\mathbf{K}\mathbf{D}^{1/2}(\mathbf{D}^{1/2}\mathbf{S})^{-1}\mathbf{M}^{1/2}$. Using eigenvalue decomposition, we can write $\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ where $\mathbf{U}$ is the matrix of orthogonal eigenvectors of $\mathbf{B}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix whose elements are the corresponding eigenvalues. Thus $Q = \mathbf{R}^T\boldsymbol{\Lambda}\mathbf{R}$, where $\mathbf{R} = \mathbf{U}^T\mathbf{M}^{-1/2}(\mathbf{Y} - \hat{\mathbf{p}})$. As nt becomes large, each element of $\mathbf{R}$ behaves very closely to a standard normal random variable and $\text{cov}(\mathbf{R}) = \mathbf{I}$. Based on this, we generate random vectors $\mathbf{R}_b^* \sim \text{MVN}(\mathbf{0}, \mathbf{I})$, $b = 1, ..., B$, and compute $Q_b^* = \mathbf{R}_b^{*T}\boldsymbol{\Lambda}\mathbf{R}_b^*$. The p-value is defined as $(1/B)\sum_{b=1}^{B} I[Q_b^* > Q]$ where I denotes the indicator function.

2.3 Simulation Study

We conducted a simulation study to evaluate the performance of our test statistic. To generate the observations we first generated threshold variables from the model

$$V_{ij} = \mathbf{X}_i^T \boldsymbol{\beta} + a h_j(\mathbf{Z}_i) + \varepsilon_{ij}, \tag{2.11}$$

where the covariates were $\mathbf{X}_i = \{1, X_{i1} \sim \mathrm{N}(0,1), X_{i2} \sim \mathrm{N}(0,1)\}^T$ and $\boldsymbol{\beta} = (1, 0.5, 0.5)^T$. We simulated $t = 3$ repeated observations for each subject $i = 1, ..., n$ with multivariate normal errors where $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ had zero mean, unit variance, and a compound symmetric correlation structure with correlation parameter θ. Then the correlated binary response was generated as $Y_{ij} = I[V_{ij} \geq 0]$. The variable a in (2.11) was used to study the size and power of our test with $a = 0$ corresponding to size and increasing values of a corresponding to power. We chose the sample size n to be 200 or 400. When $n = 200$, the choices of a were 0.25, 0.5, 0.75, 1 and when $n = 400$, $a = 0$, 0.15, 0.30, 0.45, 0.60. To simulate the SNPs, we identified the unique SNP patterns and frequency of occurrence from the SLC17A1 gene in the CATIE data (see Table 2.1). We then sampled from these unique patterns with replacement weighted by the frequency which allowed us to preserve the LD structure between the SNPs. We varied the number of SNPs by using only the 9 relevant SNPs and by adding 21 extraneous SNPs to the 9 relevant ones by sampling from $\{0, 1, 2\}$ with probability $(0.35, 0.2, 0.45)$. For the choice of h_j, we used two cases: (1) $h_1^{(1)}(\mathbf{Z}_i) = \frac{1}{4}\{2Z_{i9} - Z_{i8} + Z_{i2}^2 - (Z_{i1} - Z_{i4})^2 + Z_{i5}Z_{i6} - (Z_{i3} + Z_{i6} + Z_{i7})\}$ and $h_2^{(1)}(\mathbf{Z}_i) = h_3^{(1)}(\mathbf{Z}_i) = 0$, and (2) $h_1^{(2)}(\mathbf{Z}_i) = h_1^{(1)}(\mathbf{Z}_i) + 0.5Z_{i3}$ and $h_2^{(2)}(\mathbf{Z}_i) = h_3^{(2)}(\mathbf{Z}_i) = 0.5Z_{i3}$. Note that in Case 1, the only source of correlation is given by the parameter θ, and in Case 2, the two sources of correlation are from θ and from $h_2^{(2)}$ and $h_3^{(2)}$. We used the IBS kernel in (2.2) to model h_j. Lastly, we chose the correlation parameter of the errors to be $\theta = 0$, 0.3, 0.5, and 0.7.

Under Case 1 when $\theta = 0$, we expect the logistic KMR from Liu et al. (2008) to perform optimally because the responses are independent. Under Case 2, correlation arises from the errors and from the SNP effects and even when $\theta = 0$, the responses are not independent. We expect our proposed method to perform better for all of Case 2. We ran 20,000 simulations when $a = 0$ to examine the size of our test at $\alpha = 0.005$, 0.01, and 0.05, and 1,000 simulations to examine the power at $\alpha = 0.05$. We compared our method to the naive method of testing each of the responses $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ individually and using a Bonferroni correction for multiple testing. To perform the testing and get the p-values for the naive method, we used the `SKAT` package in R described by Wu et al. (2011).

The results comparing the size of the two methods are found in Table 2.2. Overall both

Table 2.1: Frequency table of genotype of the SLC17A1 gene in the CATIE data.

Genotype	Frequency	%
2 2 1 2 1 2 2 1 1	114	0.165
2 2 2 2 0 2 2 2 0	103	0.149
1 1 2 1 1 1 1 2 1	72	0.104
1 1 1 1 2 1 1 1 2	46	0.067
2 2 0 2 2 2 2 0 2	41	0.059
1 1 2 1 1 2 2 2 1	38	0.055
1 1 1 1 2 2 2 1 2	28	0.041
2 1 1 1 2 2 2 1 2	26	0.038
1 1 1 2 2 2 2 1 2	24	0.035
0 0 2 0 2 1 1 2 2	21	0.030
2 1 2 1 1 2 2 2 1	15	0.022
0 0 2 0 2 0 0 2 2	10	0.014
1 0 2 1 2 2 2 2 2	10	0.014
1 0 2 0 2 2 2 2 2	9	0.013
2 2 0 2 2 2 2 1 1	8	0.012
0 0 2 1 2 1 1 2 2	8	0.012
0 0 2 1 2 2 2 2 2	8	0.012
1 0 2 0 2 1 1 2 2	7	0.010
2 0 2 0 2 2 2 2 2	7	0.010
2 2 1 2 1 2 2 2 0	6	0.009
1 1 1 1 2 1 1 2 1	6	0.009
2 2 0 2 2 2 2 1 2	6	0.009
0 0 2 0 2 2 2 2 2	6	0.009
2 2 1 2 1 2 2 2 1	5	0.007
1 1 2 2 1 2 2 2 1	5	0.007
0 1 1 1 2 1 1 1 2	5	0.007
1 1 1 2 2 2 2 2 2	4	0.006
1 2 1 2 1 2 2 1 1	3	0.004
1 1 1 1 2 1 2 0 2	3	0.004
1 2 0 2 2 2 2 0 2	3	0.004
2 1 1 2 2 2 2 1 2	3	0.004
2 1 1 1 2 2 2 2 2	3	0.004
0 0 2 2 2 2 2 2 2	3	0.004
1 1 2 1 1 1 2 1 1	2	0.003
1 1 1 2 2 2 2 0 2	2	0.003
0 1 1 1 2 2 2 1 2	2	0.003
0 0 2 1 2 2 2 1 2	2	0.003
0 1 1 2 2 2 2 1 2	2	0.003
0 0 2 2 2 2 2 1 2	2	0.003
1 1 2 1 1 2 2 2 0	1	0.001
2 1 2 1 1 1 2 1 1	1	0.001
2 2 2 2 0 2 2 1 1	1	0.001
2 1 2 1 1 2 2 1 1	1	0.001
1 1 2 2 1 2 2 1 1	1	0.001
1 1 1 1 2 2 1 2 1	1	0.001
2 1 2 2 1 2 2 2 1	1	0.001
1 0 2 1 2 2 2 2 1	1	0.001
0 0 2 0 2 0 1 1 2	1	0.001
1 0 2 1 2 1 1 1 2	1	0.001
0 0 2 0 2 1 2 1 2	1	0.001
1 1 1 1 2 1 2 1 2	1	0.001
2 1 1 0 2 2 2 1 2	1	0.001
1 0 2 1 2 2 2 1 2	1	0.001
1 0 2 2 2 2 2 1 2	1	0.001
1 0 2 0 2 0 0 2 2	1	0.001
0 1 1 1 2 1 1 2 2	1	0.001
1 0 2 1 2 1 1 2 2	1	0.001
0 1 2 1 1 2 1 2 2	1	0.001
1 1 1 1 2 2 2 2 2	1	0.001
1 0 2 2 2 2 2 2 2	1	0.001

approaches had size close to the nominal values when only the relevant SNPs were included. For $n = 200$ and $\theta = 0.7$, we found that our method had inflated size for small nominal values but stabled out as the nominal value increased. For larger sample size, this inflation disappeared and our method had size very close to the nominal values. When the irrelevant SNPs were added, our method became more conservative as the sample size and θ increased. This is due to the fact that the uncorrelated irrelevant SNPs have no effect on the response and outnumber the SNPs that do. Due to the nature of the Bonferroni correction, testing the repeated observations individually became very conservative when the irrelevant SNPs were added, especially for the smaller sample size. For larger sample size, the size of the two methods were very similar for all θ.

The results for the power of the test for $n = 200$ are given in Figure 2.1. Our proposed method is denoted by "MV" for multivariate KMR and the naive approach is denoted by "UV" for univariate KMR with a Bonferroni correction. When $h_2^{(1)} = h_3^{(1)} = 0$ (top panel), the two methods generally had comparable power. For moderate to high correlations in the errors, our

Table 2.2: Type I error rates for multivariate KMR (MV) and univariate logistic KMR with Bonferroni correction (UV) for different nominal values α, sample sizes n, number of SNPs p, and correlations θ.

			$\theta = 0$			$\theta = 0.3$			$\theta = 0.5$			$\theta = 0.7$		
α			.005	.01	.05	.005	.01	.05	.005	.01	.05	.005	.01	.05
$n = 200$	$p = 9$	MV	.0058	.0114	.0556	.0052	.0112	.0561	.0057	.0113	.0515	.0074	.0121	.0548
		UV	.0054	.0108	.0508	.0055	.0107	.0499	.0052	.0103	.0489	.0051	.0104	.0460
	$p = 30$	MV	.0052	.0097	.0519	.0049	.0098	.0505	.0052	.0103	.0496	.0092	.0149	.0557
		UV	.0034	.0072	.0377	.0028	.0061	.0375	.0031	.0073	.0387	.0028	.0061	.0352
$n = 400$	$p = 9$	MV	.0063	.0110	.0515	.0053	.0103	.0511	.0051	.0101	.0492	.0054	.0094	.0442
		UV	.0061	.0108	.0511	.0053	.0114	.0502	.0056	.0107	.0471	.0057	.0104	.0474
	$p = 30$	MV	.0055	.0105	.0525	.0046	.0088	.0461	.0051	.0095	.0439	.0045	.0083	.0413
		UV	.0043	.0087	.0441	.0047	.0093	.0439	.0049	.0091	.0438	.0047	.0088	.0427

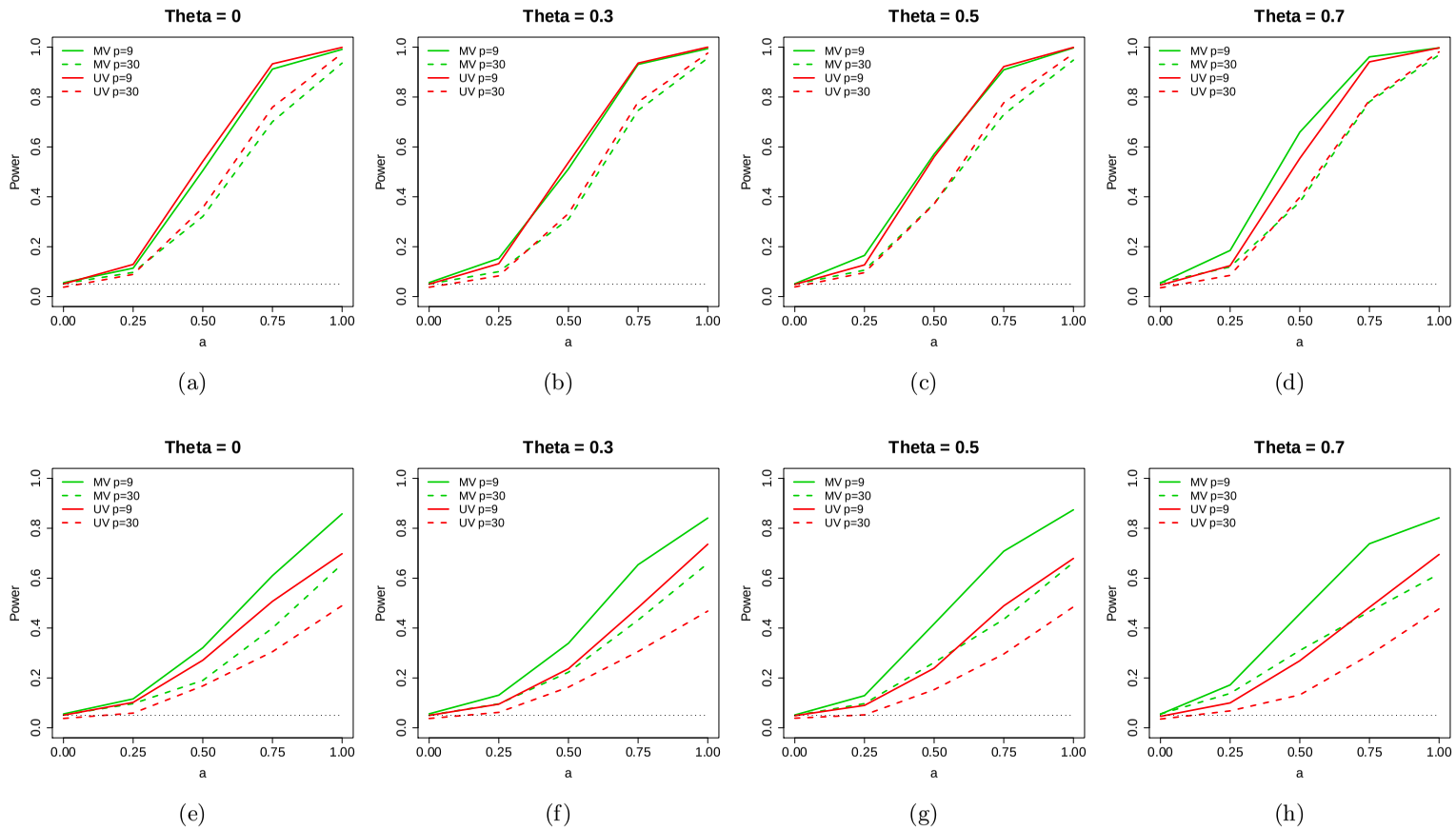

Figure 2.1: Power calculations for multivariate KMR (green lines) and univariate logistic KMR with Bonferroni correction (red lines) for sample size $n = 200$, number of SNPs $p = 9$ (solid lines) and $p = 30$ (dashed lines), and different correlations θ. The top panel (a)-(d) refers to Case 1 where the response is only correlated through the subject effects. The bottom panel (e)-(h) refers to Case 2 where the responses have correlation from shared marker effects. The effect of the SNPs is scaled by a.

method was slightly higher with a maximum gain in power of 49%. When the correlation was low or nonexistent, the naive method had higher power, but the maximum gain was only 13%. When the extraneous SNPs were included (dashed lines), we found that both methods had an overall decrease in power with both methods still very comparable. For any fixed a, as θ increased, the power of our method increased but the power of the naive method showed no discernable trend. When $h_2^{(2)}$ and $h_3^{(2)}$ were nonzero (bottom panels), we saw higher power in our proposed method for all values of a and all error correlations. This is due to the fact that our method accounts for the two sources of correlation in the responses, one from the errors and one from the SNP effects. Recall that even when $\theta = 0$ yielding uncorrelated errors, the dependence in the SNP effects is still causing the responses to be correlated. The naive method ignores this information rendering our method more powerful. For moderate to high correlations in Case 2, the increase in power using the proposed method over the naive method ranged from 21% to 74% for $p = 9$, and 29% to 135% for $p = 30$. Overall, the increase in power of the proposed method over the naive method in Case 2 is much higher than the increase of the naive method over the proposed method in Case 1. We saw an overall decrease in power for both methods when switching from Case 1 to Case 2.

The results for $n = 400$ given in Figure 2.2 follow similar trends as those for $n = 200$ in that both methods have comparable power with our method having slightly higher power for larger error correlation. For moderate to high correlation in Case 1, the maximum increase in power using our method over the naive method was 53%. For $\theta = 0$ or 0.3, the maximum increase in power using the naive method over ours was 26% corresponding to $\theta = 0$, $a = .45$, and $p = 30$. For Case 2 and when extraneous SNPs were added, we again saw our method perform universally better than the naive method. For moderate to high correlations, the increase in power using our method ranged from 29% to 133% for $p = 9$, and 39% to 87% for $p = 30$. When $\theta = 0.7$ our power's method was nearly double that of the naive method in Case 2.

In summary, our simulation study results show that both methods had appropriate size when the relevant SNPs were used, and became conservative when the irrelevant SNPs were added. We also found that the two methods are comparable when the SNP effect is sparse with our proposed method generally having slightly higher power only when θ was large. When the sample size was large and correlation is high, we did see more gain in using our method over the proposed method in terms of power. When dependency was present in the SNP effects, our proposed method had higher power than the naive method regardless of the amount of error correlation. Again, our method performs so much better because it accounts for both sources of correlation that may be present in the response. Thus when using our method over the naive

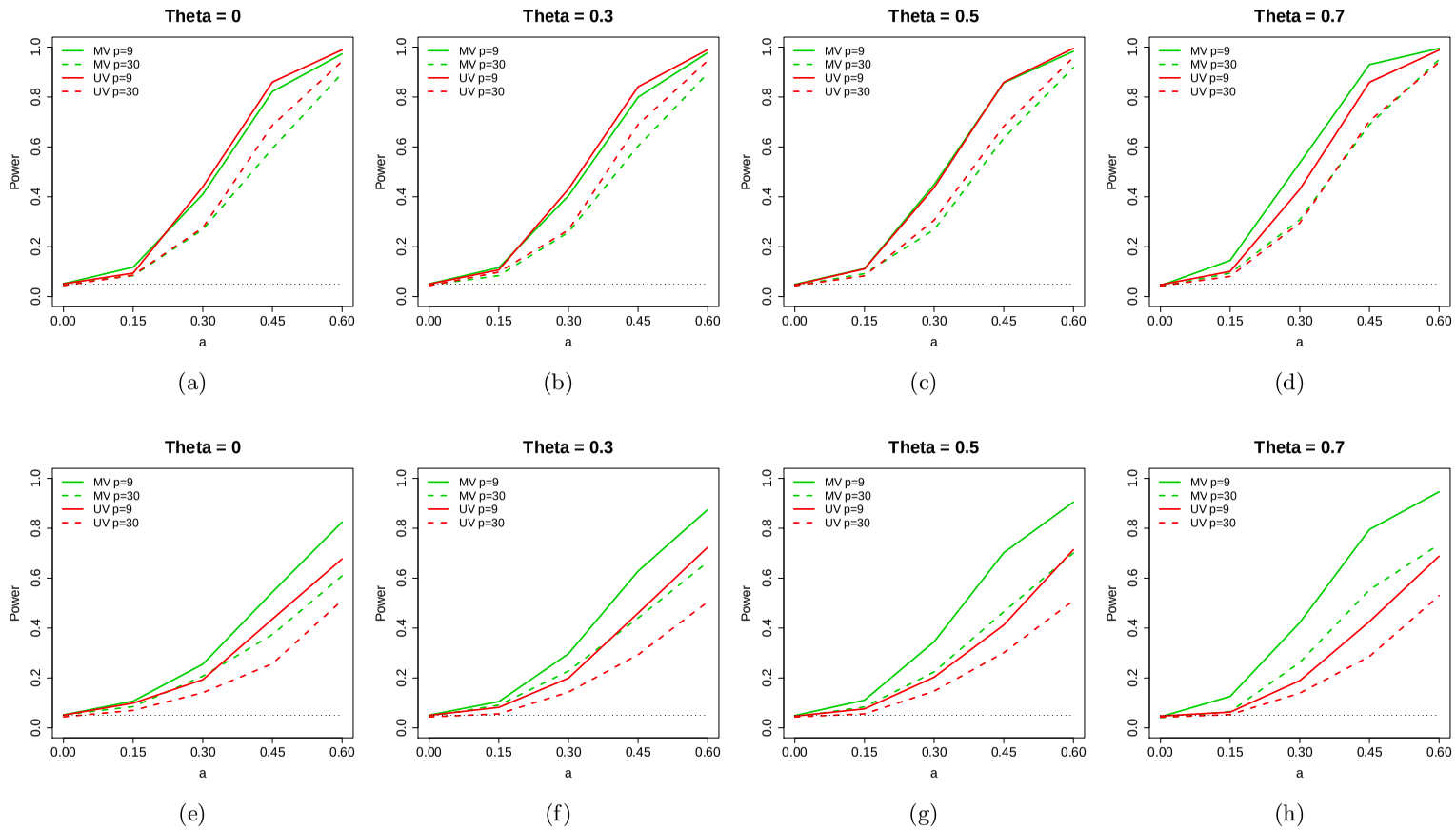

Figure 2.2: Power calculations for multivariate KMR (green lines) and univariate logistic KMR with Bonferroni correction (red lines) for sample size $n = 400$, number of SNPs $p = 9$ (solid lines) and $p = 30$ (dashed lines), and different correlations θ. The top panel (a)-(d) refers to Case 1 where the response is only correlated through the subject effects. The bottom panel (e)-(h) refers to Case 2 where the responses have correlation from shared marker effects. The effect of the SNPs is scaled by a.

method, there is no loss when correlation in the errors is small, but there is much gain if there is even a little correlation in the SNP effect no matter what the correlation in the errors is.

2.4 Data Analysis

The CATIE antibody study from Yolken et al. (2011), is based on the original CATIE study described by Lieberman et al. (2005). Studies have found that variation in the MHC region is a replicable risk factor of schizophrenia. This variation plays a known role in the body's responses to neurotrophic infectious agents. Yolken et al. (2011) also show that there is an association between exposure to these infectious agents and cognitive deficits in individuals with schizophrenia. Our main goal is to analyze the relationship between the presence or absence of antibodies of the infectious agents and the genes located in the MHC region which are associated with schizophrenia.

The philosophy of the CATIE study was to assess antipsychotic drug treatments on a wide range of individuals with schizophrenia. There were 1460 participants and 51% of them gave DNA samples with genotype data available on 492K SNPs (Sullivan et al., 2008). Using the CATIE samples, Yolken et al. (2011) measured the presence or absence of IgG class antibodies to three herpesviruses, Herpes Simplex Virus type 1 (HSV-1), Herpes Simplex Virus type 2 (HSV-2) and Cytomegalovirus (CMV). They found an association between exposure to these viruses and cognitive impairments in individuals with schizophrenia. In this analysis, we study the relationship between genes located in the MHC region that are known to be associated with schizophrenia, and presence or absence of antibodies to the herpesviruses listed above.

We applied our proposed multivariate kernel machine method to the CATIE data and compared it with the univariate method applied to each of the responses individually. The three outcomes of each individual were the presence or absence of the three herpesviruses, the covariates were sex and age, and the genetic information was the number of minor alleles at each SNP for genes associated with schizophrenia and located around the MHC region. We used the IBS kernel to model the SNP effect. We used $B = 100{,}000$ resampled statistics in order to compute the p-values for our method. For the univariate case, we reported the minimum of the three p-values after correcting for multiple testing.

The results for our data analysis are presented in Table 2.3. The multivariate and univariate methods are denoted by "MV" and "UV" respectively. Our method identified two significant genes, BTN2A2 (p-value 0.0018) and 6p22.1 (p-value 0.0015). The univariate method also identified these two regions as significant along with two additional genes, BTN2A1 and POM121L2 (both p-values 0.0012). The likely reason our method was unable to detect these two genes is

Table 2.3: Resulting p-values for multivariate (MV) and univariate (UV) testing for the CATIE Antibody study using the IBS kernel. Each gene had a different sample size, n, and different number of SNPs p. The Bonferroni correction is $0.05/12 = 0.0042$ in order to analyze 12 genes and significant results are in bold.

Gene	n	p	MV	UV
SLC17A1	690	9	0.0074	0.0042
SLC17A3	388	13	0.4452	0.2430
BTN3A2	632	5	0.1070	0.0534
BTN2A2	419	5	**0.0018**	**0.0006**
BTN2A1	661	4	0.0044	**0.0012**
HIST1H2AG	419	2	0.5366	0.3387
HIST1H2BJ	645	2	0.1180	0.0135
PRSS16	408	1	0.2739	0.1257
POM121L2	643	4	0.0405	**0.0012**
ZNF184	389	8	0.9296	1
NOTCH4	587	24	0.2118	0.1356
6p22.1	490	787	**0.0015**	**<0.0001**

because there were a small number of SNPs in the set with little correlation, similar to Case 1 in our simulation where the UV method was more favorable. Another reason could be that the relationship among the SNPs is not well captured by the IBS kernel. When the linear and quadratic kernels were used (not presented), our method was able to detect these and other regions as significant. Choosing the correct kernel to use is a recurring problem in the KM literature since in practice, the true interaction between SNPs in a set is unknown.

2.5 Discussion

In this chapter, we model and test the joint effect of a marker set on multiple binary responses using kernel machine regression. The kernel method allows for nonlinearity and complicated interactions among markers in the same pathway and avoids the issues of parametric techniques and assumptions. Our procedure adjusts for other covariates and accounts for two sources of correlation that can arise in multiple outcomes, namely the random individual effect, and the shared marker effects. We developed a score-based test using the GEE framework to determine the joint marker effects on all the responses collectively. We conducted a simulation study to assess the performance of our method compared to evaluating the responses individually and using a Bonferroni correction. We found that the two methods were comparable when the responses were independent or had weak correlation solely due to the individual random effect. When correlation from shared markers was introduced, our method was uniformly more powerful power, regardless of individual random effect. Because there is no meaningful loss when

correlation is low and considerable gain when correlation is moderate to high, we recommend our method as a general tool for multiple binary phenotype analysis.

In this work, we focus on SNP-set effects using the IBS kernel, but our method is not restricted to just genetic data. Environmental predictors and any other types of factors of interest that have non-linear effects, complicated interactions, and even interactions with markers can be modeled in this framework as long as an appropriate kernel is chosen. One could then develop tests to determine if these factors have significant effect on the binary traits or if they significantly interact with markers. Thus our method can be used as a device in studying phenotypic data with genetic predictors or general correlated binary outcomes with environmental factors.

Our method can also serve as a preliminary analysis in a genome-wide association study (GWAS). In GWAS, many markers across the entire genome are identified as important to the trait or phenotype of interest. As mentioned in the introduction, testing marker-by-marker has many drawbacks, and testing all makers together will have very little, if any power to detect signals. Thus marker sets can be appropriately established gene-by-gene and our methodology can be used as a screening process to identify important genes by detecting significant global effects of the marker sets. Once these genes are identified, one can conduct further gene-level analysis to interpret the global signal and to determine which SNPs in the set are driving the global effect. We note that our method is computationally efficient and scalable to genome wide studies. For $n = 100$ individuals, $t = 3$ responses per individual, $p = 30$ SNPs, $B = 10000$, and using the IBS kernel, on average, our test takes 27.7 seconds to run on an Intel i7 2.93 GHz processor (using one core) with 8 GB of RAM.

Chapter 3

Parametrically Guided Estimation in Varying Coefficient Models

3.1 Introduction

A common scenario in studies is when the researcher observes some covariates and a response and wants to estimate the conditional mean function of the response given the covariates. A typical approach is to fit a generalized linear model (GLM) (Nelder and Wedderburn, 1972) where a parametric assumption is made on a transformation of the conditional mean. This approach is easily interpreted and efficient if the correct parametric model is chosen, but can have serious consequences when the model is misspecified, a common problem in real-world applications. Alternatively, nonparametric methods make little or no assumptions of the model and are robust to model misspecification, however, they are slower to converge (Glad, 1998) and can fail when the dimensions of the covariates are too high (Fan and Zhang, 2008). In this chapter, we combine the benefits of both parametric and nonparametric methods by considering estimation of nonparametric varying coefficient models using a pre-specified parametric family of functions as a guide. We use a local likelihood kernel smoothing based estimation procedure and investigate the asymptotic properties of the resulting estimates. We evaluate the finite sample performance of our method via a simulation study and demonstrate it by applying to the AIDS Clinical Trials Group (ACTG) Protocol 315 data described in Section 3.5.

Varying coefficient models (VCMs) are nonparametric generalized additive models (Wood, 2006) that increase the flexibility of linear models by using a smooth function to model the parameters. There are numerous advantages of using VCMs over their parametric linear model counterparts. First, fitting a standard linear model is too restrictive because few real world problems satisfy the assumption of linearity. The true, possibly nonlinear, relationship between the

covariates are not well captured by polynomial fitting which can lead to large bias in estimation. Second, VCMs allow for the interaction between covariates to be modeled in a nonparametric way. Models with only main effects included may miss the effect from the interaction between covariates. Another advantage for using VCMs is that the dimensionality issue is avoided because the coefficient functions are allowed to vary as a function of another covariate (Hastie and Tibshirani, 1993). VCMs are easy to interpret and arise when it is desirable to know how regression coefficients change over groups, or over time in longitudinal studies. These flexible models can be applied to a variety of data types, including longitudinal data (Brumback and Rice, 1998; Hoover et al., 1998), time series data (Chen and Tsay, 1993; Huang and Shen, 2004), environmental data (Fan and Zhang, 1999), and genetic data (Ma et al., 2011).

VCMs consist of smooth, functional parameters that need to be estimated and this can be done using penalized spline approaches (Eilers and Marx, 1996; Cao et al., 2010), basis expansion methods (Holdeman, 1969), or by applying regression locally (Chen and Tsay, 1993; Cai et al., 2000). In this chapter, we utilize the latter method since these estimators are efficient and have nice sampling properties (Fan, 1993). This method involves using a kernel function to weight the likelihood and applying polynomial regression locally using Taylor series approximations. In practice, the full likelihood may be unknown or difficult to construct, so we replace it with the quasi-likelihood introduced by Wedderburn (1974). With the quasi-likelihood, only the relationship between the mean and the variance needs to be specified, and the model will still retain most of the efficiency of a maximum likelihood estimation procedure.

Nonparametric estimators can be enhanced by using a parametric guide. In practice, previous information or exploratory analysis may give some insight on the shape of the unknown functions and this information can be used to speed up convergence and reduce bias. The basic process is as follows: 1) identify a parametric family of functions that captures the shape, 2) remove the trend and carry out local polynomial estimation, and 3) add the trend back to obtain the final estimators of the functions. This guided estimation scheme was first studied in the density estimation framework where Hjort and Glad (1995) showed that their estimator had better bias and similar variance compared to the traditional nonparametric estimator, even when the guide was superficial. It has also been studied in least squares regression (Glad, 1998; Martins-Filho et al., 2008) quasi-likelihood models (Fan et al., 2009) and nonparametric additive models (Fan et al., 2013).

In this chapter we use a parametric guide to enhance local polynomial estimators, and thus extend the work of Fan et al. (2009) and Fan et al. (2013) to VCMs. We use quasi-likelihoods as a general case when the likelihood is unavailable. We borrow the idea of parametrically guided estimation to improve the bias of our estimators and we develop a new optimal bandwidth selection methodology for the kernel weight function when we apply local polynomial regression.

We show in simulations that the guided estimators have lower bias and similar variance when a fixed bandwidth is used, and lower bias and variance when the optimal bandwidth is used. We then estimate the functional parameters in the (ACTG) Protocol 315 data.

The rest of this chapter is organized as follows. in Section 3.2 we give an overview of QLMs and the standard nonparametric estimation procedure using local polynomial fitting. We propose our parametrically guided estimation scheme in Section 3.2.2 using two different types of corrections, and give some asymptotic properties of our estimators. In Section 3.3, we present one method of choosing the bandwidth parameter in local polynomial fitting. We evaluate the performance of our estimators compared to the standard ones in Section 3.4 and found that our estimators had lower bias when a fixed bandwidth was used, and lower bias and variability when the optimal bandwidth was chosen. We then applied our methodology to the ACTG data in Section 3.5 and provide some concluding remarks in Section 3.6.

3.2 Guided Estimation for Varying Coefficient Models

Assume that for each of n subjects we observe covariates $\mathbf{X}_i = (1, X_{i1}, ..., X_{iq})^T$ and T_i, and a response Y_i. The VCM for these covariates is defined as

$$\begin{aligned} g(\mu_i) &= \theta_0(T_i) + X_{i1}\theta_1(T_i) + \cdots + X_{iq}\theta_q(T_i) \\ &= \mathbf{X}_i^T \boldsymbol{\theta}(T_i), \end{aligned} \tag{3.1}$$

where $\mu_i = E(Y_i|\mathbf{X}_i, T_i)$ is the conditional mean of the response, $g(\cdot)$ is a link function from the GLM framework, and $\boldsymbol{\theta}(\cdot) = \{\theta_0(\cdot), \theta_1(\cdot), ..., \theta_q(\cdot)\}^T$ are unknown, smooth functions. The first term models the unique effect of T and the remaining terms model the interaction between $\mathbf{X}$ and T. This VCM is more flexible than a linear regression model because it allows the effect of $\mathbf{X}$ to vary smoothly with T and the effect of T is not restricted to a linear assumption. The goal is to estimate $\boldsymbol{\theta}(\cdot)$ and there are several ways to do this (Cleveland et al., 1991; Hastie and Tibshirani, 1993). The method we adopt is to use local-likelihood kernel smoothing using a quasi-likelihood.

3.2.1 Framework and Local Likelihood Estimation

Quasi-likelihood models (QLMs), an extension of GLMs, are ideal because often the full likelihood may be unknown or difficult to construct. In QLMs, only the relationship between the conditional mean and variance of the responses need to be specified, which is often doable in practice. The full conditional log-likelihood is replaced with a quasi-likelihood function $Q(Y_i, \mu_i)$

and if we define $\text{var}(Y_i|\mathbf{X}_i, T_i) = V(\mu_i)$, then Q satisfies

$$\frac{\partial}{\partial \mu} Q(y, \mu) = \frac{y - \mu}{V(\mu)}.$$

Wedderburn (1974) shows that Q has similar properties to the log-likelihoods and that Q is exactly the likelihood when the response comes from a single parameter exponential family.

The standard, unguided, nonparametric procedure for estimating $\boldsymbol{\theta}(\cdot)$ is to use local polynomial fitting by first approximating the functions using a Taylor series expansion so that

$$\theta_j(T_i) \approx \beta_{j0} + (T_i - t_0)\beta_{j1} + \cdots + (T_i - t_0)^P \beta_{jP}, \tag{3.2}$$

where $\beta_{jp} = \theta_j^{(p)}(t_0)/p!$ for $p = 0, ..., P$ and $j = 0, ..., q$. Substituting this approximation into (3.1) yields $g(\mu_i) = \mathbf{G}_i^T \boldsymbol{\beta}$ where $\mathbf{G}_i = \{1, (T_i - t_0)^1, ..., (T_i - t_0)^P\}^T \otimes \mathbf{X}_i$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, ..., \boldsymbol{\beta}_P^T)^T$ where $\boldsymbol{\beta}_p = (\beta_{0p}, ..., \beta_{qp})^T$. The approximation in (3.2) is only accurate when t_0 is close to T_i, so the quasi-likelihood is weighted in such a way that more weight is given to t_0's close to T_i and little to no weight to those far from T_i. This is done by using a kernel function $K_h(\cdot) = K(\cdot/h)/h$ and defining the local quasi-likelihood as

$$\sum_{i=1}^{n} Q\{g^{-1}(\mathbf{G}_i^T \boldsymbol{\beta}), Y_i\} K_h(T_i - t_0). \tag{3.3}$$

The parameter h is the bandwidth and needs to be estimated (see Section 3.3). We maximize (3.3) with respect to $\boldsymbol{\beta}$ and the solution $\hat{\boldsymbol{\beta}}_0$ will be the estimate of $\boldsymbol{\theta}(t_0)$.

3.2.2 Guided Estimation

The local likelihood estimators can be enhanced by using a parametric guide. Intuitively speaking, more curvature in a function makes it more difficult to estimate. If, through exploratory analysis or prior information, one has some idea of the shape of the true function, then one can identify a parametric family that captures this trend. Using the parametric guide, the curvature of the function can be removed yielding a flatter curve that is easier to estimate. Once this flatter curve is estimated, then the guide can be used to add the trend back and obtain the final estimate of the original function. This type of guided estimation has been shown to reduce bias of nonparametric estimators (Hjort and Glad, 1995; Glad, 1998; Martins-Filho et al., 2008) and improve variance since a larger bandwidth can be selected (Fan et al., 2009, 2013).

Define a parametric family that captures the trend of the function as $\{\theta_{jg}(t, \boldsymbol{\alpha}_j) \colon \boldsymbol{\alpha}_j = (\alpha_{j1}, ..., \alpha_{jm_j})^T \in \mathbb{A} \subset \mathbb{R}^{m_j}\}$ for $j = 0, ..., q$. The optimal guides can be found by maximizing

the quasi-likelihood

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{\mathbf{X}_i^T \boldsymbol{\theta}_g(T_i, \boldsymbol{\alpha})\right\}, Y_i\right]$$

with respect to $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, ..., \boldsymbol{\alpha}_q^T)^T$ where $\boldsymbol{\theta}_g(T_i, \boldsymbol{\alpha}) = \{\theta_{0g}(T_i, \boldsymbol{\alpha}_0), ..., \theta_{qg}(T_i, \boldsymbol{\alpha}_q)\}^T$. The best fit is denoted by $\hat{\boldsymbol{\theta}}_g(t) = \boldsymbol{\theta}_g(t, \hat{\boldsymbol{\alpha}})$ where we suppress the dependence on $\boldsymbol{\alpha}$ in our notation. In this section we present two methods of removing the trend using an additive correction or a multiplicative correction.

Additive Correction

If the curvature of θ_j is well approximated by $\hat{\theta}_{jg}$, then estimating the quantity $\theta_j(t) - \hat{\theta}_{jg}(t)$ will yield more accurate and less variable estimates since this quantity is close to flat. Once estimated, the guide is added back to give the final estimate of θ_j. This process can be achieved in one step by defining $\eta_j(t) = \theta_j(t) - \hat{\theta}_{jg}(t) + \hat{\theta}_{jg}(t_0)$ and estimating η_j at t_0. This definition of η_j is known as the additive correction.

Using this correction, the VCM in (3.1) can be rewritten as

$$g(\mu_i) = \mathbf{X}_i^T \boldsymbol{\eta}(T_i) + \mathbf{X}_i^T \mathbf{h}(T_i), \tag{3.4}$$

where $\boldsymbol{\eta}(t) = \{\eta_0(t), ..., \eta_q(t)\}^T$, and $\mathbf{h}(t) = \{\hat{\theta}_{0g}(t) - \hat{\theta}_{0g}(t_0), ..., \hat{\theta}_{qg}(t) - \hat{\theta}_{qg}(t_0)\}^T$. Similar to Section 3.2.1 and with a slight abuse of notation, a Taylor series approximation is used for $\eta_j(T_i)$ about the point t_0 such that

$$\eta_j(T_i) \approx \beta_{j0} + (T_i - t_0)\beta_{j1} + \cdots + (T_i - t_0)^P \beta_{jP},$$

where $\beta_{jp} = \eta_j^{(p)}(t_0)/p!$ for $p = 0, ..., P$. The local quasi-likelihood

$$\begin{aligned} Q(\boldsymbol{\beta}) &\equiv Q(\boldsymbol{\beta}; h, t_0, \hat{\boldsymbol{\alpha}}) \\ &= \sum_{i=1}^{n} Q\left[g^{-1}\left\{\mathbf{G}_i^T \boldsymbol{\beta} + \mathbf{X}_i^T \mathbf{h}(T_i)\right\}, Y_i\right] \times K_h(T_i - t_0) \end{aligned}$$

is maximized with respect to $\boldsymbol{\beta}$ and the estimate of $\hat{\boldsymbol{\beta}}_0$ corresponds to $\hat{\boldsymbol{\eta}}(t_0)$ which gives the final estimate $\hat{\boldsymbol{\theta}}(t_0)$. Because $\mathbf{h}(t)$ is known for fixed $T = t$, our model can be fit using standard software with $\mathbf{h}(t)$ as an offset.

Multiplicative Correction

An alternative correction which leads to a different guided estimator is the multiplicative correction. As in the additive case, the ratio $\theta_j(t)/\hat{\theta}_{jg}(t)$ will be flat if $\hat{\theta}_{jg}(t)$ captures the trend of $\theta_j(t)$ and estimating this ratio will be less biased than estimating the unknown function directly. Once estimated, the ratio is then multiplied by the guide to get the final estimate of θ_j. The multiplicative correction is defined as $\eta_j(t) = \theta_j(t)\{\hat{\theta}_{jg}(t_0)/\hat{\theta}_{jg}(t)\}$ and the one step solution requires estimating $\eta_j(t)$ at t_0.

Using the multiplicative correction, (3.1) is written as

$$g(\mu_i) = \frac{\hat{\theta}_{0g}(T_i)}{\hat{\theta}_{0g}(t_0)}\eta_0(T_i) + \frac{\hat{\theta}_{1g}(T_i)}{\hat{\theta}_{1g}(t_0)}X_{i1}\eta_1(T_i) + \cdots + \frac{\hat{\theta}_{qg}(T_i)}{\hat{\theta}_{qg}(t_0)}X_{iq}\eta_q(T_i).$$

Estimating $\eta_j(T)$ is achieved by first using a Taylor series expansion of $\eta_j(t)$ about the point t_0 and then maximizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q\left\{g^{-1}(\mathbf{G}_i^{*T}\boldsymbol{\beta}), Y_i\right\} K_h(T_i - t_0)$$

with respect to $\boldsymbol{\beta}$ where $\mathbf{G}_i^* = \{1, T_i - t_0, ..., (T_i - t_0)^p\} \otimes \left(\frac{\hat{\theta}_{0g}(T_i)}{\hat{\theta}_{0g}(t_0)}, X_{i1}\frac{\hat{\theta}_{1g}(T_i)}{\hat{\theta}_{1g}(t_0)}, ..., X_{iq}\frac{\hat{\theta}_{qg}(T_i)}{\hat{\theta}_{qg}(t_0)}\right)$. The solution $\hat{\boldsymbol{\beta}}_0$ give our final estimates of $\hat{\boldsymbol{\theta}}(t_0)$. By manipulating the design matrix, there is no offset for the multiplicative correction and this model can easily be fit using standard GLM software.

3.2.3 Asymptotic Properties

In this section, we give an overview of the asymptotic properties of the guided estimators. Define $\kappa_d = \int u^d K(u)\,du$ and $\nu_d = \int u^d K^2(u)\,du$. Define $\mathbf{M}$ to be a 2×2 matrix with elements $\mathbf{M}_{kl} = \kappa_{k+l-2}$ and $\mathbf{R}$ to be a 2×2 matrix with elements $\mathbf{R}_{kl} = \nu_{k+l-2}$. Let $\rho(\mathbf{x}, t) = 1/[V\{\mu(\mathbf{x}, t)\}g'^2\{\mu(\mathbf{x}, t)\}]$ and $\gamma(\mathbf{x}, t_0) = \text{var}(Y_1|\mathbf{X}_1 = \mathbf{x}, T_1 = t_0)/[V\{\mu(\mathbf{x}, t_0)\}g'\{\mu(\mathbf{x}, t_0)\}]^2$. Make the definitions

$$\begin{aligned}
\mathbf{V}_1(t_0) &= f_t(t_0)\mathbf{M} \otimes E\{\rho(\mathbf{X}_1, T_1)\mathbf{X}_1\mathbf{X}_1^T|T_1 = t_0\},\\
\mathbf{V}_2(t_0) &= f_t(t_0)\mathbf{R} \otimes E\{\gamma(\mathbf{X}_1, T_1)\mathbf{X}_1\mathbf{X}_1^T|T_1 = t_0\},\\
\mathbf{B}_1(t_0) &= \mathbf{M}^{-1}(\kappa_2, \kappa_3)^T \otimes \boldsymbol{\eta}''(t_0)/2,
\end{aligned}$$

where $f_t(\cdot)$ is the marginal density function of T. Then for a fixed guide, we have the following result.

Theorem 1. *Fix a point t_0 and assume the guide is fixed. Under certain regulatory conditions, as $h \to 0$, $nh \to \infty$, and $nh^5 \to$ constant, we have*

$$(nh)^{1/2}\{\hat{\boldsymbol{\theta}}(t_0) - \boldsymbol{\theta}(t_0) - h^2 \boldsymbol{B}_1(t_0)\} \to^d \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is the leading $(q+1) \times (q+1)$ submatrix of $\boldsymbol{V}_1^{-1}\boldsymbol{V}_2\boldsymbol{V}_1^{-1}$.

If there is no (or constant) guide and the model belongs to a one-parameter exponential family with the canonical link and correctly specified variance function, then the results in Theorem 1 are the same as those in Theorem 1 of Cai et al. (2000). This theorem assumes a fixed guide but it can be shown that the expressions above are still valid for estimated guides as well. For further details and proof of Theorem 1, see Davenport et al. (2013). In general, when the bandwidth is fixed and an appropriate guide is chosen, then the squared bias of our estimators are reduced. Furthermore, finding the optimal bandwidth using the procedure from Section 3.3 allows for a larger bandwidth to be selected compared to the standard nonparametric estimates which will reduce the variance as well as the bias. We demonstrate this in our simulation study in Section 3.4.

3.3 Optimal Bandwidth Selection

Note that for simplicity of presentation, we will consider our bandwidth selection method using the additive correction; the multiplicative correction follows easily by using the multiplicative definition of $\eta_j(\cdot)$, replacing $\mathbf{G}_i$ with $\mathbf{G}_i^*$, and omitting the offset term $\mathbf{X}_i^T\mathbf{h}(T_i)$.

Once $\hat{\boldsymbol{\beta}}$ is obtained, the bias arises from the approximation term of the Taylor series expansions. Hence, using more terms in the series should theoretically produce less bias. Let $r_j(T_i) = \eta_j(T_i) - \sum_{k=0}^{P} \eta_j^{(k)}(t_0)(T_i - t_0)^k/k!$ be the approximation error. If a higher order Taylor approximation is substituted for $\eta_j(T_i)$ in r_j, then $r_j(T_i) \approx \sum_{k=1}^{a} \eta_j^{(p+k)}(t_0)(T_i - t_0)^{p+k}/(p+k)! \stackrel{\text{def}}{=} r_{ji}$. We then maximize the local quasi-likelihood including the approximation errors

$$Q^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q\left[g^{-1}\left\{\mathbf{G}_i^T\boldsymbol{\beta} + \mathbf{X}_i^T\mathbf{h}(T_i) + \mathbf{X}_i^T\mathbf{r}_i\right\}, Y_i\right] \times K_h(T_i - t_0)$$

with respect to $\boldsymbol{\beta}$, where $\mathbf{r}_i = (r_{0i}, ..., r_{qi})^T$. Define the maximizer as $\hat{\boldsymbol{\beta}}^*$. The local quasi-likelihood $Q^*(\boldsymbol{\beta})$ is differentiated with respect to $\boldsymbol{\beta}$ to get the gradient vector

$$Q^{*\prime}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Y_i - g^{-1}\{\mathbf{G}_i^T\boldsymbol{\beta} + \mathbf{X}_i^T\mathbf{h}(T_i) + \mathbf{X}_i^T\mathbf{r}_i\}}{V\left[g^{-1}\{\mathbf{G}_i^T\boldsymbol{\beta} + \mathbf{X}_i^T\mathbf{h}(T_i) + \mathbf{X}_i^T\mathbf{r}_i\}\right]} \left(g^{-1}\right)' (\mathbf{G}_i^T\boldsymbol{\beta} + \mathbf{X}_i^T\mathbf{h}(T_i) + \mathbf{X}_i^T\mathbf{r}_i)\mathbf{G}_i K_h(T_i - t_0)$$

and the second derivative is taken to get the Hessian matrix $Q^{*\prime\prime}(\boldsymbol{\beta})$. A Taylor series expansion is then applied to $Q^{*\prime}$ around $\hat{\boldsymbol{\beta}}$ to get

$$Q^{*\prime}(\hat{\boldsymbol{\beta}}^*) \approx Q^{*\prime}(\hat{\boldsymbol{\beta}}) + Q^{*\prime\prime}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = \mathbf{0}$$

and thus, an approximation of the estimation bias is $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^* \approx \{Q^{*\prime\prime}(\hat{\boldsymbol{\beta}})\}^{-1} Q^{*\prime}(\hat{\boldsymbol{\beta}})$.

To get an approximation of the variance, a Taylor series expansion of $Q'(\hat{\boldsymbol{\beta}})$ is done about the true $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}^0$. Note that

$$\mathbf{0} = Q'(\hat{\boldsymbol{\beta}}) \approx Q'(\boldsymbol{\beta}^0) + Q''(\boldsymbol{\beta}^0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0),$$

which implies $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \approx -\{Q''(\boldsymbol{\beta}^0)\}^{-1} Q'(\boldsymbol{\beta}^0)$, and the estimate of the conditional variance is

$$\mathrm{var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 | \mathbf{X}_i, T_i) \approx \{Q''(\boldsymbol{\beta}^0)\}^{-1} \mathrm{var}\{Q'(\boldsymbol{\beta}^0) | \mathbf{X}_i, T_i\} \{Q''(\boldsymbol{\beta}^0)\}^{-1},$$

where $Q''(\boldsymbol{\beta}^0)$ can be approximated by $Q''(\hat{\boldsymbol{\beta}})$. To approximate the variance term, note that

$$\begin{aligned}
\mathrm{var}\{Q'(\boldsymbol{\beta}^0)|\mathbf{X}_i, T_i\} &= \sum_{i=1}^{n} \mathrm{var}\left(\frac{\partial}{\partial \boldsymbol{\beta}} Q\left[g^{-1}\left\{\mathbf{G}_i^T \boldsymbol{\beta} + \mathbf{X}_i^T \mathbf{h}(T_i)\right\}, Y_i\right] \middle| \mathbf{X}_i, T_i \right) K_h^2(T_i - t_0) \\
&= \sum_{i=1}^{n} \mathrm{var}\left(\frac{Y_i - g^{-1}\{\mathbf{G}_i^T \boldsymbol{\beta} + \mathbf{X}_i^T \mathbf{h}(T_i)\}}{V\left[g^{-1}\{\mathbf{G}_i^T \boldsymbol{\beta} + \mathbf{X}_i^T \mathbf{h}(T_i)\}\right]} (g^{-1})'\{\mathbf{G}_i^T \boldsymbol{\beta} + \mathbf{X}_i^T \mathbf{h}(T_i)\} \middle| \mathbf{X}_i, T_i \right) \\
&\quad \times \mathbf{G}_i \mathbf{G}_i^T K_h^2(T_i - t_0) \\
&\approx \frac{\left[(g^{-1})'\{\mathbf{X}_i^T \boldsymbol{\theta}(t_0)\}\right]^2}{V\left[g^{-1}\{\mathbf{X}_i^T \boldsymbol{\theta}(t_0)\}\right]} \mathbf{G}_i \mathbf{G}_i^T K_h^2(T_i - t_0).
\end{aligned}$$

The last approximation follows from the fact that T_i only has significant weight in the neighborhood of t_0.

To compute the MSE, we denote the bias of $\hat{\boldsymbol{\theta}}$ as $\mathbf{B}(t_0; h) = \{B_0(t_0; h), ..., B_q(t_0; h)\}^T$ corresponding to the first $q+1$ components of $[Q^{*\prime\prime}(\hat{\boldsymbol{\beta}})]^{-1} Q^{*\prime}(\hat{\boldsymbol{\beta}})$. The variance-covariance matrix $\mathbf{V}(t_0; h)$ of $\hat{\boldsymbol{\theta}}$ is the first $(q+1) \times (q+1)$ submatrix of $\mathrm{var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 | \mathbf{x}, t)$ and the variance of the estimated VCM $\mathbf{X}_i^T \hat{\boldsymbol{\theta}}$ is $\mathbf{X}_i^T \mathbf{V}(t_0; h) \mathbf{X}_i$. The conditional MSE of $\mathbf{X}^T \hat{\boldsymbol{\theta}}$ given $\mathbf{X} = \mathbf{x}$ is

$$\mathrm{MSE}(t_0; h) = \mathbf{x}^T \left\{ \mathbf{B}(t_0; h) \mathbf{B}(t_0; h)^T + \mathbf{V}(t_0; h) \right\} \mathbf{x}.$$

The sample MSE is derived as

$$\begin{aligned}\widehat{\mathrm{MSE}}(t_0;h) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\left\{\mathbf{B}(t_0;h)\mathbf{B}(t_0;h)^T+\mathbf{V}(t_0;h)\right\}\mathbf{X}_i \\ &= \mathrm{trace}\left[\left\{\mathbf{B}(t_0;h)\mathbf{B}(t_0;h)^T+\mathbf{V}(t_0;h)\right\}n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\right].\end{aligned}$$

We propose to choose h such that

$$\hat{h} = \mathrm{argmin}_h \int \widehat{\mathrm{MSE}}(t;h)\,dt.$$

To summarize, we use a grid of t_0's and a grid of candidate bandwidths. We fit the local quasi-likelihood for each t_0 and h candidate, calculate the ERSC from Fan et al. (1998), then sum the ERSCs over t_0. The bandwidth with the lowest sum becomes the pilot bandwidth. Using this bandwidth, we fit a local quasi-likelihood for each t_0 to obtain $\hat{\boldsymbol{\beta}}$. For a new set of candidate bandwidths, we fit the local quasi-likelihood using higher order Taylor series approximation and obtain $\hat{\boldsymbol{\beta}}^*$, which is theoretically more accurate. We compute the bias and variance using the gradient and Hessian of the quasi-likelihood, compute the MSE and the candidate bandwidth with the lowest MSE is our optimal bandwidth.

3.4 Simulation

We conducted a simulation study to evaluate the performance of our estimators. We generated each observation (X_i, T_i, Y_i) by first simulating the covariates X_i and T_i from a uniform distribution. Then, given X_i and T_i, the conditional mean μ_i of the response was generated as

$$\mu_i = g^{-1}\left\{\theta_0(T_i) + X_{i1}\theta_1(T_i)\right\},$$

where g is the canonical link. We used a grid of equally spaced values t_k for $k = 1, ..., K = 100$ to estimate the two functions. A cubic guide $\theta_{0g}(t, \boldsymbol{\alpha}_0) = \alpha_{01} + \alpha_{02}t + \alpha_{03}t^2 + \alpha_{04}t^3$ was used for estimating θ_0 and a quadratic guide $\theta_{1g}(t, \boldsymbol{\alpha}_1) = \alpha_{11} + \alpha_{12}t + \alpha_{13}t^2$ was used for estimating θ_1. We used local linear polynomial estimators with the Epanechnikov kernel weight, and for bias calculations, we chose the degree of the Taylor expansion to be $a = 1$.

For $R = 1000$ simulations, we generated the data and estimated the parameters for the cubic and quadratic guides as described above. To get the final estimates $\hat{\theta}_0$ and $\hat{\theta}_1$, we maximized the local quasi-likelihood with the appropriate distribution and canonical link, and Epanechnikov kernel weight. The design matrix of the local likelihood was constructed using both the addi-

tive and multiplicative corrections ($\mathbf{G}_i$ and $\mathbf{G}_i^*$, respectively). To find the optimal smoothing bandwidth, we first simulated 15 datasets and applied our methods from Section 3.3. We took the median of these 15 values as the optimal bandwidth and used that value as fixed for the 1000 simulations. We also compared the two methods using the optimal bandwidth from the unguided method as the fixed bandwidth.

Once the two functions were estimated, we computed the marginal squared bias, marginal variance, and marginal MSE of each. Define $B_{jk} = R^{-1}\sum_{r=1}^{R}[\hat{\theta}_{j,r}(t_k) - \theta_j(t_k)]$, $V_{jk} = R^{-1}\sum_{r=1}^{R}[\hat{\theta}_{j,r}(t_k) - R^{-1}\sum_{r'=1}^{R}\hat{\theta}_{j,r'}(t_k)]^2$, and $\text{MSE}_{jk} = B_{jk}^2 + V_{jk}$ where $j = 0, 1$ and r indexes the simulation. The average marginal squared bias of $\hat{\theta}_j$ is $B_j^2 = K^{-1}\sum_{k=1}^{K} B_{1k}^2$, the average marginal variance is $V_j = K^{-1}\sum_{k=1}^{K} V_{jk}$ and the average marginal MSE is $\text{MSE}_j = K^{-1}\sum_{k=1}^{K}\text{MSE}_{jk}$. However instead of averaging over all k, we used the 10% trimmed mean. Tables 3.1 - 3.3 show the results of our simulations. "Opt h" is each estimation method's optimal smoothing bandwidth obtained via the bias-variance tradeoff in Section 3.3. All values for squared bias, variance, and MSE in the table are multiplied by 100. "Fix h" refers to the fixed bandwidth obtain from the optimal bandwidth from the unguided estimators.

Example 1: Poisson Response

For the Poisson response, $n = 100$ or 200 covariates X_i and T_i were generated with $X_i \sim \text{Unif}[-1, 1]$ and $T_i \sim \text{Unif}[-2, 2]$. The true functions were $\theta_0(t) = \sin(\pi t/2) + 4$ and $\theta_1(t) = \sin(\pi t/4 - \pi/2)/2 + 1$. The response Y_i was generated from a Poisson(μ_i). We used a grid of $K = 100$ equally spaced values in $[-2, 2]$ for t_0 to estimate the two functions. Table 3.1 gives the (trimmed) average marginal squared bias, variance, and MSE for the two functions estimated by the original method using no guide and by our method using guided estimation with additive and multiplicative corrections. When the same bandwidth is used the guided estimation procedure reduces bias but has no effect on variance. When the optimal bandwidth is used, the guided estimates have lower bias and lower variance. This is because the guides account for much of the trend in the true curve and the nonparametric correction is flatter and easier to estimate, resulting in lower bias. As the sample size increased, we saw a reduction in bias, variance and MSE.

Example 2: Normal Response

For the Gaussian response, $n = 100$ or 200 covariates X_i and T_i were generated with $X_i \sim \text{Unif}[-1, 1]$ and $T_i \sim \text{Unif}[-2, 2]$. The true functions were $\theta_0(t) = \sin(\pi t/2) - 2$ and $\theta_1(t) = 2\sin(\pi t/4 - \pi/2) + 3$. The response Y_i was generated from a Normal($\mu_i, 1$). Table 3.2 gives the (trimmed) average marginal squared bias, variance, and MSE for the two functions estimated

Table 3.1: Results of trimmed average bias, variance, and MSE for the unguided estimators (Naive), and the guided estimators with the additive (Add) and multiplicative (Mul) corrections for Example 1. "Fix" refers to the estimates using a fixed bandwidth and "Opt" refers those using the optimal bandwidth. All values are multiplied by 100.

			Bias^2			Variance			MSE		
n	h		Naive	Add	Mul	Naive	Add	Mul	Naive	Add	Mul
100	Fix	$\hat{\theta}_0(t)$	0.129	0.022	0.022	0.179	0.173	0.174	0.308	0.195	0.196
		$\hat{\theta}_1(t)$	0.003	0.001	0.001	0.552	0.522	0.522	0.556	0.522	0.522
	Opt	$\hat{\theta}_0(t)$	0.129	0.098	0.096	0.179	0.104	0.106	0.308	0.203	0.202
		$\hat{\theta}_1(t)$	0.003	0.001	0.001	0.552	0.294	0.296	0.556	0.295	0.297
200	Fix	$\hat{\theta}_0(t)$	0.080	0.014	0.015	0.090	0.088	0.089	0.171	0.103	0.104
		$\hat{\theta}_1(t)$	0.003	0.001	0.001	0.269	0.261	0.262	0.272	0.262	0.262
	Opt	$\hat{\theta}_0(t)$	0.080	0.026	0.026	0.090	0.074	0.075	0.171	0.100	0.101
		$\hat{\theta}_1(t)$	0.003	0.001	0.001	0.269	0.217	0.217	0.272	0.217	0.217

Table 3.2: Results of trimmed average bias, variance, and MSE for Example 2. All values are multiplied by 100.

			Bias^2			Variance			MSE		
n	h		Naive	Add	Mul	Naive	Add	Mul	Naive	Add	Mul
100	Fix	$\hat{\theta}_0(t)$	1.05	0.11	0.10	3.82	4.15	4.29	4.86	4.26	4.39
		$\hat{\theta}_1(t)$	0.20	0.02	0.02	11.68	11.81	12.21	11.89	11.83	12.23
	Opt	$\hat{\theta}_0(t)$	1.05	0.22	0.21	3.82	3.41	3.51	4.86	3.63	3.73
		$\hat{\theta}_1(t)$	0.20	0.04	0.03	11.68	8.73	9.92	11.89	8.77	8.96
200	Fix	$\hat{\theta}_0(t)$	0.46	0.07	0.07	2.20	2.31	2.38	2.66	2.38	2.45
		$\hat{\theta}_1(t)$	0.13	0.01	0.01	6.65	6.70	6.85	6.78	6.71	6.85
	Opt	$\hat{\theta}_0(t)$	0.46	0.11	0.11	2.20	2.00	2.05	2.66	2.11	2.16
		$\hat{\theta}_1(t)$	0.13	0.01	0.01	6.65	5.55	5.63	6.78	5.56	5.64

Table 3.3: Results of trimmed average bias, variance, and MSE for Example 3 for $n = 500$. All values are multiplied by 100.

		Bias^2		Variance		MSE	
		Naive	Additive	Naive	Additive	Naive	Additive
Fix h	$\hat{\theta}_0(t)$	0.84	0.17	62.43	67.23	63.27	67.40
	$\hat{\theta}_1(t)$	0.21	0.02	27.10	29.08	27.31	29.10
Opt h	$\hat{\theta}_0(t)$	0.84	0.18	62.34	57.60	63.27	57.77
	$\hat{\theta}_1(t)$	0.21	0.02	27.10	24.95	27.31	24.96

by the original method using no guide and by our method using guided estimation with additive and multiplicative corrections. In this example, the guided estimates still have lower bias and variance when the optimal bandwidth is used. When the same bandwidth is used, the variance of the additive and multiplicative correction is slightly higher than the unguided estimates. The gains in bias reduction are counteracted by the higher variance so the MSE is approximately the same for both methods.

Example 3: Bernoulli Response

For the Bernoulli response, we chose a larger sample size of $n = 500$ since the estimation of the success probability is more difficult than estimating the mean in the Gaussian and Poisson case. The covariates X_i and T_i were generated with $X_i \sim \text{Unif}[1, 2]$ and $T_i \sim \text{Unif}[-1, 1]$. The true functions were $\theta_0(t) = \sin(\pi t)/2 + 1$ and $\theta_1(t) = 0.7 \sin\{(t + 1)\pi/2\} - 1$. The response Y_i was generated from a Bernoulli(μ_i). The results from this example are presented in Table 3.3. Using a multiplicative correction with Bernoulli data is very unstable due to the possibility of dividing by zero, thus this correction was not used. We see that using the guides reduces bias and variance when the optimal bandwidth is used, and reduces bias but has little effect on variance when a fixed bandwidth is used.

3.5 HIV Data Analysis

In the AIDS Clinical Trials Group (ACTG) Protocol 315, 48 individuals infected with HIV-1 were given potent antiviral medicine to evaluate the efficacy of treatment on the reduction of viral load (plasma HIV-1 RNA). The viral load was measured repeatedly over three months and 31 baseline covariates were measured for each individual. Details of this study can be found in Lederman et al. (1998) and Liang et al. (2003). Of these 31 covariates, Wu and Wu (2002) identified those that were significant predictors and in this illustration, we chose two of these covariates that appeared frequently in their models, baseline viral load and baseline CD4+

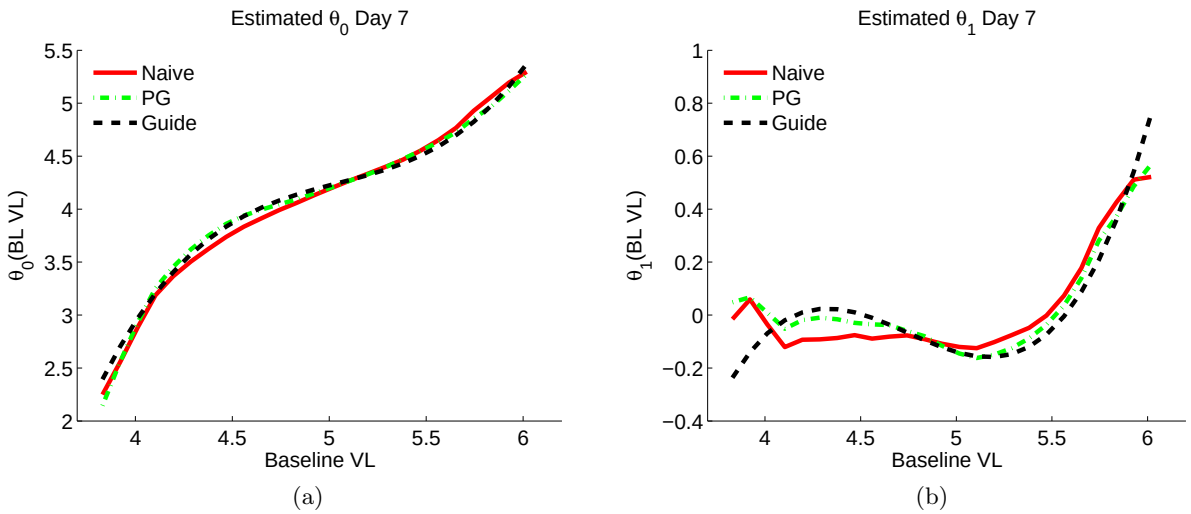

Figure 3.1: Nonparametric estimate (solid red) and parametrically guided estimate (green dot-dashed) of θ_0 (a) and θ_1 (b) along with the cubic guide (black dashed).

counts. Viral load is the number of copies per milliliter and is measured on a logarithmic scale with base 10. CD4+ counts are the number of lymphocytes that are CD4+. The response was the change from baseline viral load measured at day 7. If an individual did not have a viral load measurement on day 7, then the preceding or following day measurement was used. We would like to determine if baseline viral load and baseline CD4+ counts have an effect on the change from baseline viral load measurement while adjusting for interaction between them.

Using a grid of size of 25 for t_0 and the Epanechnikov kernel, we fit the model in (3.1) and estimated θ_0 and θ_1 using the original unguided nonparametric method, represented by the solid red line in Figure 3.1. The pre-asymptotic bandwidth selector gave bandwidth 0.67. Based on the shape of these unguided estimates, we chose a cubic guide for our guided estimation of θ_0 and θ_1 and used the additive correction. The results from our fit are represented by the green dot-dashed lines in Figure 3.1 and the guides used are the black dashed line. The bandwidth for our method was also 0.67. Our estimated $\hat{\theta}_0$ had more curvature than the naive counterpart and followed the parametric guide very closely, suggesting that there is little model misspecification when using a cubic guide for θ_0. For $\hat{\theta}_1$, the naive estimate and our guided estimate had somewhat similar shapes, with the most difference in the left endpoints.

The pointwise bootstrap 95% confidence intervals are given in Figure 3.2. The confidence intervals for the guided estimates were slightly wider in the boundaries but overall were similar to those of the naive estimate. In Figure 3.2(c) and (d), the entire confidence interval for the functions contain zero. Recall that θ_1 is the slope function and this term models the interaction

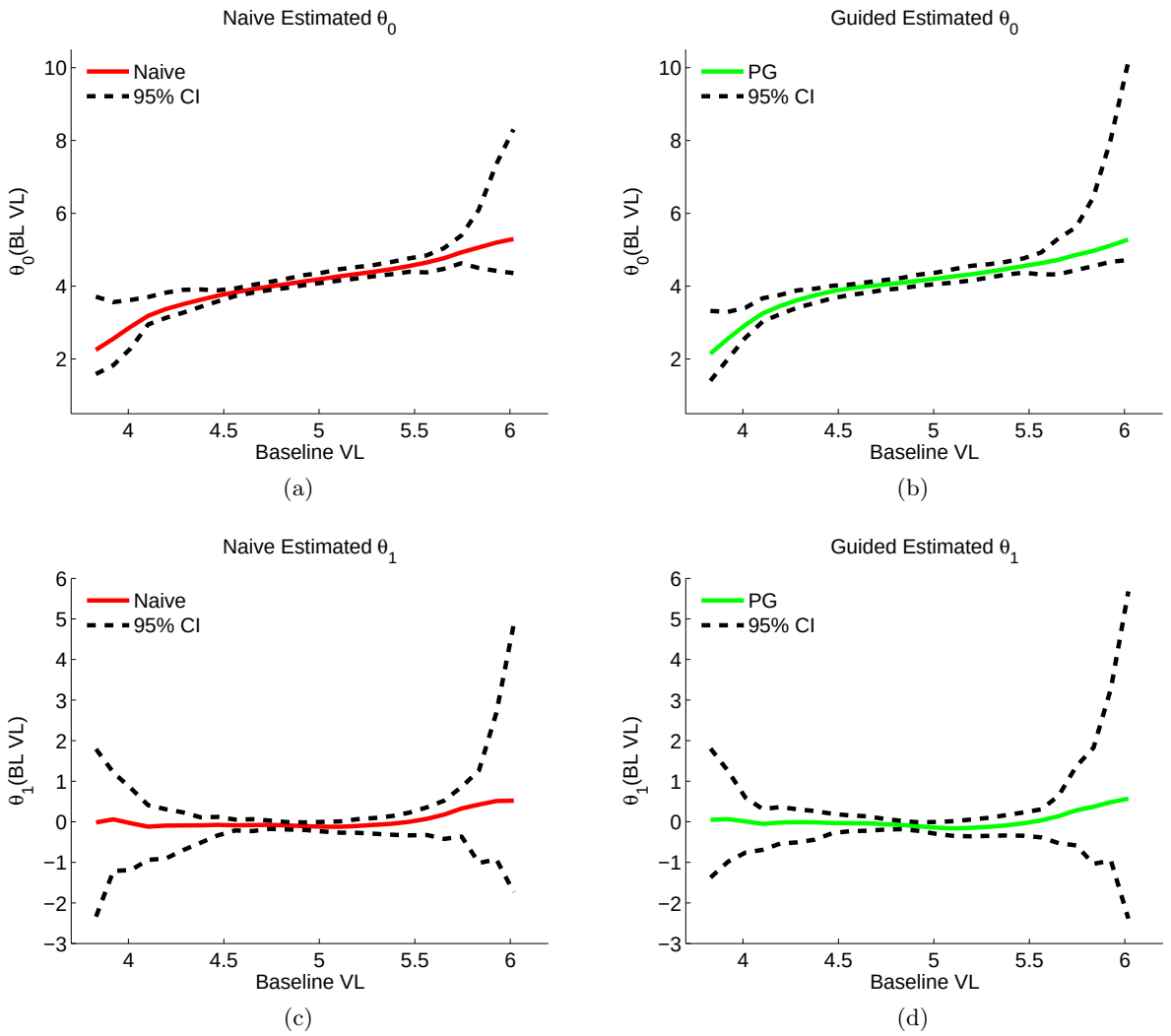

Figure 3.2: Nonparametric estimate (solid red) of θ_0 (a) and θ_1 (c) and parametrically guided estimate (solid green) of θ_1 (b) and θ_2 (d) along with a pointwise 95% bootstrap confidence intervals (black dashed).

between the two covariates. This indicates that there is no interaction between baseline viral load and baseline CD4+ counts and the response can be adequately modeled as a cubic function of baseline viral load alone.

We also used the VCM to separately model the viral load after day 14, day 21, and day 28 with the same two covariates in order to compare the estimated functions on different days to day 7. Again, if an individual did not have a viral load measurement on these exact days, then the preceding or following day was used, and if all three days were missing, the individual was

dropped from the analysis. This yielded 35 individuals for the day 14 analysis, 39 individuals for the day 21 analysis, and 38 individuals for the day 28 analysis. The estimated functions corresponding to these responses are given in Figure 3.3. The pre-asymptotic bandwidth selector gave bandwidths 0.53, 1.22 and 0.77 for the naive estimates of day 14, 21, and 28 respectively. The bandwidth for the guided estimates were 0.63, 1.26, and 0.77 for day 14, 21, and 28 respectively. The shape of the slope function θ_1 changes drastically for the different days, but the entire confidence interval for all three θ_1 functions (not presented) contains zero. Thus CD4+ has a very different interaction effect on baseline viral load for different days, but the overall effect is not significant. The shape of the intercept function θ_0 was similar for days 14 and 21, and had more of a cubic shape for day 28. Similar to day 7, the parametrically guided estimates of θ_0 followed their respective guides very closely.

3.6 Discussion

In this chapter, we used parametric guides to enhance the performance of nonparametric estimators of the parameter functions in varying coefficient models. We generalized to quasi-likelihoods since the true likelihood is often unavailable. We presented two ways of using the guide and estimated the corrected functions using local polynomial fitting. We developed the asymptotic properties of the guided estimators and a method of selecting the optimal bandwidth parameter of the kernel function. We conducted a simulation study to compare our guided estimators to their standard nonparametric equivalents, and found that the guided estimators had lower bias when a fixed bandwidth was used, and lower bias and variance when the optimal bandwidth was used. In general, even if the shape of the parameter function is not captured by the guide, the guided estimator will still have better bias than the unguided counterpart and the two will have similar variability.

In this chapter we present the additive and multiplicative correction which are special cases of a unified family of guided estimators proposed by Fan et al. (2009). This work could be extended to include this unified family and its asymptotic properties. Other future work includes extending our methods to functional data where the covariates of interest are smooth functions and the response can be functional or scalar. The functional covariates correspond to unknown functional parameters that need to be estimated, and this can be done using our guided estimation scheme. This work can also be extended to multivariate unknown functions (e.g. $\theta(\mathbf{T}_i)$) by using an empirical basis expansion of the function and reducing it to a VCM.

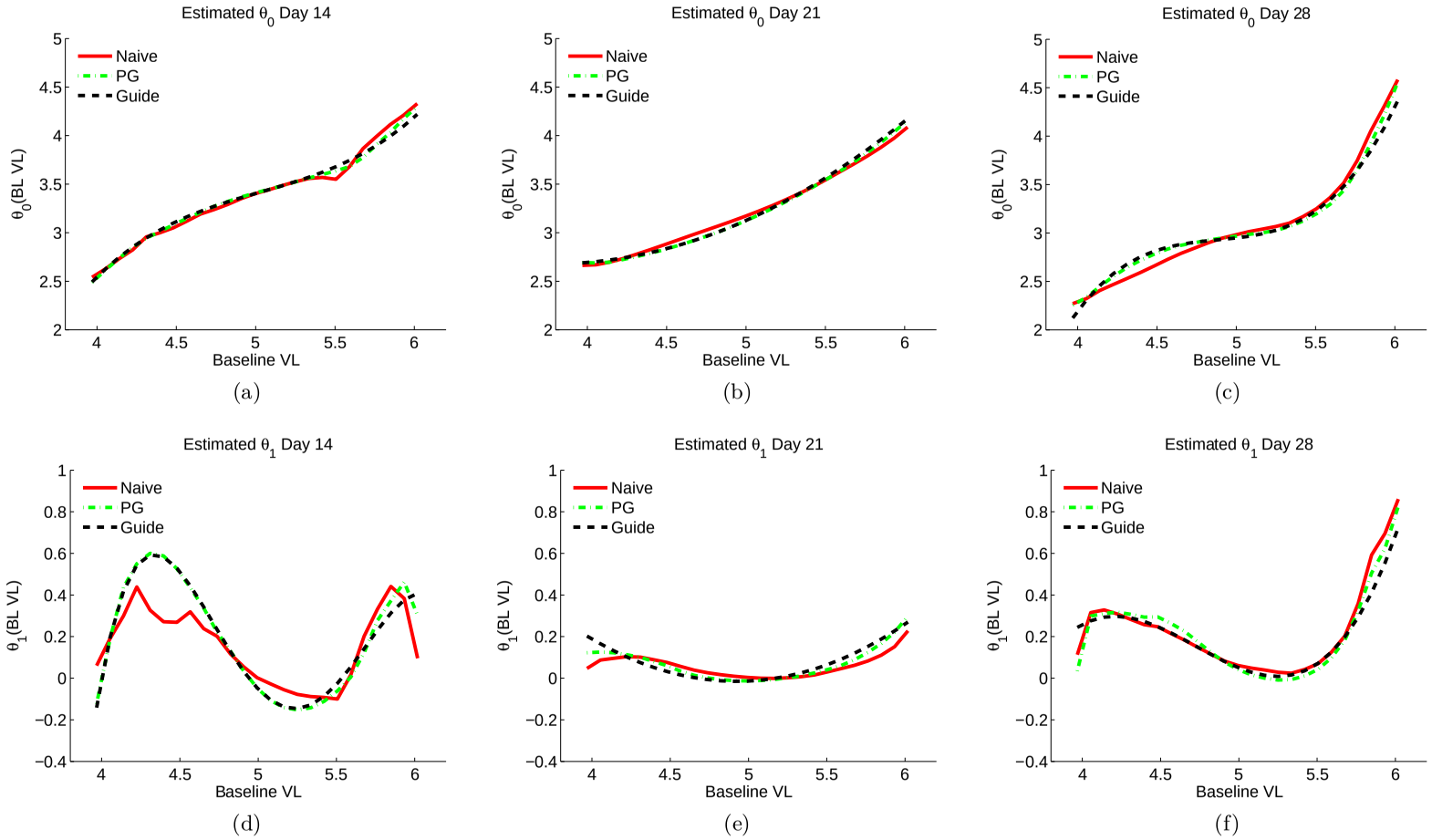

Figure 3.3: Nonparametric estimate (solid red) and parametrically guided estimate (green dot-dashed) of θ_0 (a)-(c) and θ_1 (d)-(f) for day 14, day 21, and day 28 responses. A cubic guide was used in (a), (c), and (f), a quadratic guide was used in (b) and (e), and a quartic guide was in (d).

Chapter 4

Modeling Interaction using Functional Regression Models

4.1 Introduction

Functional linear regression is an extension of familiar regression analysis to the case where one or more predictors are functions, rather than vector valued. These covariates arise when multiple measurements are taken on a subject and these measurements can be thought of as discrete realizations of an unknown, smooth, underlying process. Functional data has been studied in a wide variety of areas including environmental health (Meiring, 2007; Torres et al., 2011; Crambes et al., 2009), entomology (Carey et al., 1998), genetics (Baladandayuthapani et al., 2010; Wu and Müller, 2010), biology and medicine (Crainiceanu et al., 2009; Ullah and Finch, 2010; Staicu et al., 2012), geophysics (Maslova et al., 2010) and finance (Laukaitis, 2008; Gabrys et al., 2010). See Ullah and Finch (2013) for many more applications dealing with functional data.

In a typical scenario, we observe realizations of a functional covariate and a scalar response and interest lies in determining the relationship between the two. One of the most straightforward and widely used tools to achieve this goal is the functional linear model (Müller and Stadtmüller, 2005; Cardot, 2007; Crambes et al., 2009; Ferraty et al., 2012). Analogous to the generalized linear model for vector valued predictors, the functional linear model quantifies the effect of the functional predictor by the inner product between the predictor and an unknown functional parameter. The goal is to make inference on the functional parameter, and this can be complicated by measurement error or missing observations in the functional covariate. Overall functional linear models are simple and easy to interpret, but may be too restrictive in

capturing the true effects.

In this chapter, we focus on the case where a functional covariate is measured along with a scalar covariate of interest with the goal of determining the relationship between the predictors and a scalar response. This is challenged by the fact that the functional covariate may be noisy and sparsely observed, and that it may interact with the scalar covariate. The functional linear model can account for the noise in the data, but can not capture the interaction effect. Thus, our objective is to accurately model the effects of the scalar covariate and functional covariate of interest, while accounting for the interaction between the two, and adjusting for extraneous covariates that may have an effect on the response. We do this by modeling the main effect of the scalar covariate nonparametrically using a smooth parameter function, and by modeling the main effect of the functional covariate jointly with the interaction effect using another bivariate parameter function. We model the extraneous covariates that do not interact with the predictors in a parametric way. The ultimate goal is to make inference on these model components.

There are a number of models and estimation procedures for fitting functional and scalar data. Cardot and Sarda (2008) present a flexible model that we adopt in this paper, however they only present the case for Gaussian responses, they do not propose a way to obtain the standard errors of their estimators, and they assume that the entire functional covariate is observed without error, which is not practical in real world applications. Wu et al. (2010) fit a similar model for Gaussian responses using a different estimation scheme, and they do account for noisy or irregularly spaced observations in the functional covariate. However, they do not compute standard errors either. Both Cardot and Sarda (2008) and Wu et al. (2010) use local estimation procedures that are computationally intense.

We propose a new varying-coefficient functional linear model, and present a general model using two new estimation procedures that can easily be extended to generalized responses and noisy or sparsely observed functional covariates. The simplicity of our proposed methodology allows us to easily find the standard errors of our estimators and thus, construct confidence bands and surfaces and develop testing procedures. The global nature of our estimators makes them computationally inexpensive and less variable than those found by local estimation schemes.

The rest of this chapter is organized as follows. In Section 4.2 we present a varying coefficient model and propose a scheme to estimate the parameters of the model. In Section 4.3, we propose two estimation methods for a more flexible model and suggest a naive testing procedure in Section 4.4. We show how our methodology is easily extended to more general settings in Section 4.5. We conduct a simulation study in Section 4.6, apply our methods to real data in Section 4.7, and give concluding remarks in Section 4.8.

4.2 Linear Varying coefficient Model in the Functional Framework

We begin by presenting the extension of the varying coefficient model into the functional framework, then present the general model in Section 4.3.

Assume that for each of n subjects we observe a scalar response Y, demographic covariates $\mathbf{W}$, a noisy functional covariate Z, and another scalar covariate G that can interact with Z. For ease of presentation, we assume that the response, Y_i comes from a Gaussian distribution conditional on the covariates. Extension to the generalized case where Y_i comes from an exponential family is straightforward and is presented in Section 4.5. The demographic covariates, such as age and sex, are stored in a $q \times 1$ vector $\mathbf{W}_i$. The functional covariate is observed on a dense grid of points $\{t_m\}_{m=1}^M$ within a closed interval for each individual and these measurements can be thought of as discrete realizations of an unknown smooth curve, $X_i(t)$. In practice, the functional covariate is measured with error, and thus instead of observing the true $X_i(t)$'s, we actually observe $Z_i(t) = X_i(t) + \delta_i(t)$, where $\delta(t)$ is a white noise process with mean 0 and variance σ_δ^2. The grid $\{t_m\}_{m=1}^M$ is assumed to be the same for each subject (such as all subjects being measured in 30-minute intervals) but there can be missing values. Finally, there is a scalar covariate of interest G_i that can possibly interact with the functional covariate $X_i(t)$. The goal is to study the effects of G and $X(t)$ on the response Y while accounting for the possible interaction between these two covariates and adjusting for the effect of the other covariates measured.

Recall from Chapter 3 that varying coefficient models (VCMs) are a type of generalized additive models (GAMs) (Wood, 2006) that allows coefficients to vary as a function of another covariate (Hastie and Tibshirani, 1993). Consider the following model, a VCM in the functional framework,

$$Y_i = \mathbf{W}_i^T \boldsymbol{\gamma} + \beta_0(G_i) + \int_{t \in \mathcal{T}} X_i(t)\beta_1(t)\, dt + G_i \int_{t \in \mathcal{T}} X_i(t)\beta_2(t)\, dt + \varepsilon_i, \tag{4.1}$$

where we assume $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. The demographic covariates are modeled parametrically with corresponding parameter vector $\boldsymbol{\gamma}$ and the main effect of the interacting covariate G is modeled nonparametrically by an unknown smooth function $\beta_0(\cdot)$. The first integral term models the main effect of the functional covariate X and the last term models the interaction between G and X. The unknown smooth functions $\beta_1(\cdot)$ and $\beta_2(\cdot)$ are the corresponding parameter functions. We would like to fit (4.1) by estimating the parameters $\boldsymbol{\gamma}$, $\beta_0(\cdot)$, $\beta_1(\cdot)$, and $\beta_2(\cdot)$. This functional linear VCM is a natural model to use when G is dichotomous.

4.2.1 Functional Principal Component Analysis

In practice, the true $X(t)$'s are unknown and we only observe discrete realizations of these curves. The observed curves may not be smooth for a variety of reasons, such as the realizations are measured with error, there is an insufficient number of measurements, or some measurements are missing. Before we can fit the VCM in (4.1) we estimate each of the smooth unobserved curves by $\hat{X}_i(t)$ for $i = 1, ..., n$ using functional principal component analysis (FPCA). There are several smoothing approaches in the FPCA literature (Staniswalis and Lee, 1998; Yao et al., 2005; Di et al., 2009; Goldsmith et al., 2013) and the one we adopt in this chapter is the same as Di et al. (2009) and Goldsmith et al. (2013). FPCA is a dimension reduction technique that decomposes observed curves so that the most important modes of variation in the curve can be identified and highlights the characteristics of the data (Ramsay and Silverman, 2005; Yao et al., 2005).

Assume that $X_i(t)$ is square integrable with mean function $\theta(t)$. FPCA is stepwise procedure where the first step involves finding the weight function $\phi_1(t)$ such that $\xi_{i1} = \int \phi_1(t)\{X_i(t) - \theta(t)\}\, dt$ has the largest variance subject to $\|\phi_1\|^2 = 1$. The constraint is needed so that the problem is well defined. The second step is to find the weight function $\phi_2(t)$ such that $\xi_{i2} = \int \phi_2(t)\{X_i(t) - \theta(t)\}\, dt$ has the largest variance subject to $\|\phi_2\|^2 = 1$ and ϕ_2 is orthogonal to ϕ_1. The first step identifies the largest and most important mode of variation while the second and subsequent steps identify important modes of variation that are different from the previous steps. This process continues until a sufficient number of weight functions is chosen, typically based on percent of variability explained by the combined steps.

FPCA is motivated by the fact that it finds the set of orthonormal basis functions that approximates the true underlying curves as close as possible. Finding the principal components of a curve is the same as finding the eigenvalues and eigenfunctions of the covariance function. Denote the covariance function of $X_i(t)$ as $K_i(s,t) = \text{cov}\{X_i(s), X_i(t)\}$. By Mercer's theorem (Cristianini and Shawe-Taylor, 2000), $K_i(\cdot,\cdot)$ can be represented by an orthonormal basis of eigenfunctions and nonnegative eigenvalues so that $K_i(s,t) = \sum_{p=1}^{\infty} \lambda_{ip}\phi_p(s)\phi_p(t)$, where the λ_{ip}'s are the ordered eigenvalues and the ϕ_p's are the corresponding eigenfunctions. We can then represent $X_i(t)$ in terms of the truncated Karhunen-Loève expansion

$$X_i(t) = \theta(t) + \xi_{i1}\phi_1(t) + \xi_{i2}\phi_2(t) + \cdots + \xi_{iP}\phi_P(t).$$

The scores are $\xi_{ip} = \int \phi_p(t)\{X_i(t) - \theta(t)\}\, dt$ and are uncorrelated (Di et al., 2009).

When the realizations are measured with error, we actually observe $Z_i(t) = X_i(t) + \delta_i(t)$

where $\delta_i(t)$ is a white noise process with variance σ_δ^2. Define

$$C_i(s,t) = \text{cov}\{Z_i(s), Z_i(t)\} = \text{cov}\{X_i(s), X_i(t)\} + \sigma_\delta^2 I(t = s),$$

where I is the indicator function. The noisy curves are smoothed by first estimating the mean function $\theta(t)$ using some type of smoother and denote this estimate $\hat{\theta}(t)$. Using this estimate, define $G_i(s,t) = \{Z_i(s) - \hat{\theta}(t)\}\{Z_i(t) - \hat{\theta}(t)\}$ as the rough covariances. We then smooth these covariances for $\{G_i(s,t) \colon s \neq t\}$ using a two-dimensional smoother, call this $\hat{G}(s,t)$, and then predict the values for $\hat{G}(s,t)$ when $s = t$. Lastly, we estimate the error variance as $\hat{\sigma}_\delta^2 = \int \hat{C}_i(t,t) - \hat{G}_i(t,t)\,\mathrm{d}t$. The scores can be found by doing an eigenvalue decomposition of $\hat{G}_i(s,t)$ and can be truncated to a certain number P by using some criteria such as the percent of variability explained by them. The final curves are reconstructed as

$$\hat{X}_i(t) = \hat{\theta}(t) + \hat{\xi}_{i1}\hat{\phi}_1(t) + \hat{\xi}_{i2}\hat{\phi}_2(t) + \cdots + \hat{\xi}_{iP}\hat{\phi}_P(t).$$

See Ramsay and Silverman (2005) or Di et al. (2009) for a more in depth review of FPCA.

4.2.2 Estimation of Parameters

Once $\hat{X}_i(t)$ is obtained, the unknown parameters in the model can be estimated. We use basis expansions to represent the unknown parameter functions. That is

$$\beta_0(g) = \sum_{j=1}^{J} \eta_{0j}\psi_{0j}(g), \quad \beta_1(t) = \sum_{k=1}^{K} \eta_{1k}\psi_{1k}(t), \quad \text{and} \quad \beta_2(t) = \sum_{l=1}^{L} \eta_{2l}\psi_{2l}(t), \tag{4.2}$$

where ψ_{0j}, ψ_{1k}, and ψ_{2l} are B-spline basis functions and η_{0j}, η_{1k}, and η_{2l} are unknown parameter coefficients. A common concern in the functional data literature is how to determine the number of basis functions J, K, and L and how to choose the actual basis functions. Ramsay and Silverman (2005) suggest the general rule of using a Fourier basis for cyclic data and a B-spline basis for data without a periodic trend. The choice of the basis function could also depend on derivative estimation. When using a B-spline basis, order and knot selection become a factor. These selections can be arbitrary or data driven. See Ramsay and Silverman (2005) for a review of different basis functions and their drawbacks. In this chapter, we use B-splines for our basis expansions, but this can easily be replaced with any choice of basis.

Using (4.2) and the estimated $\hat{X}_i(t)$, (4.1) can be rewritten as

$$\begin{aligned} Y_i &= \mathbf{W}_i^T\boldsymbol{\gamma} + \sum_{j=1}^{J}\eta_{0j}\psi_{0j}(G_i) + \int \hat{X}_i(t)\sum_{k=1}^{K}\eta_{1k}\psi_{1k}(t)\,dt + G_i\int \hat{X}_i(t)\sum_{l=1}^{L}\eta_{2l}\psi_{2l}(t)\,dt + \varepsilon_i \\ &= \mathbf{W}_i^T\boldsymbol{\gamma} + \boldsymbol{\psi}_{i0}^T\boldsymbol{\eta}_0 + \mathbf{c}_{i1}^T\boldsymbol{\eta}_1 + G_i\mathbf{c}_{i2}^T\boldsymbol{\eta}_2 + \varepsilon_i, \end{aligned} \tag{4.3}$$

where $\boldsymbol{\psi}_{i0} = \{\psi_{01}(G_i), ..., \psi_{0J}(G_i)\}^T$, $\boldsymbol{\eta}_0 = (\eta_{01}, ..., \eta_{0J})^T$, $\mathbf{c}_{i1}$ is a $K \times 1$ vector with elements $\int \hat{X}_i(t)\psi_{1k}(t)\,dt$ for $k = 1, ..., K$, $\mathbf{c}_{i2}$ is a similar $L \times 1$ vector with elements $\int \hat{X}_i(t)\psi_{2l}(t)\,dt$ for $l = 1, ..., L$, $\boldsymbol{\eta}_1 = (\eta_{11}, ..., \eta_{1K})^T$, and $\boldsymbol{\eta}_2 = (\eta_{21}, ..., \eta_{2L})^T$. This is a linear model with $q + J + K + L$ parameters, which, in practice, can be large number. For example, if we use an order four B-spline basis with six interior knots for all three parameter functions, then $J + K + L = 30$.

Fitting standard linear model techniques will likely result in estimates that are undersmooth. The smoothness of the parameter functions can be controlled by adding a penalty to the objective, usually the squared second derivative of the function (Wood, 2006). Thus the penalized negative log-likelihood is

$$\sum_{i=1}^{n}\ell(Y_i; \mathbf{W}_i, G_i, \hat{X}_i) + \lambda_0\int \left\{\beta_0''(g)\right\}^2 dg + \lambda_1\int \left\{\beta_1''(t)\right\}^2 dt + \lambda_2\int \left\{\beta_2''(t)\right\}^2 dt,$$

where $\ell(Y_i; \mathbf{W}_i, G_i, \hat{X}_i) = (Y_i - \mathbf{W}_i^T\boldsymbol{\gamma} - \boldsymbol{\psi}_{i0}^T\boldsymbol{\eta}_0 - \mathbf{c}_{i1}^T\boldsymbol{\eta}_1 - G_i\mathbf{c}_{i2}^T\boldsymbol{\eta}_2)^2$ and λ_0, λ_1, and λ_2 are smoothing parameters that control the tradeoff between model fit and model complexity. For $m = 0, 1, 2$, when $\lambda_m = 0$, the estimate of $\beta_m(\cdot)$ is unpenalized, and when $\lambda_m \to \infty$, the estimate of $\beta_m(\cdot)$ tends to a straight line. The integrated squared second derivative is a common way to penalize estimates that are too rough (Ramsay and Silverman, 2005) and so we adopt it here.

Because we are using (4.2) to approximate the parameter functions, it can be shown that $\int \{\beta_m''(v)\}^2\,dv = \boldsymbol{\eta}_m^T\mathbf{S}_m\boldsymbol{\eta}_m$ where the (i, j)th element of $\mathbf{S}_m$ is $\int \psi_{mi}''(v)\psi_{mj}''(v)\,dt$. Thus, similar to O'sullivan et al. (1986) and Gu and Kim (2002), we minimized the penalized negative log-likelihood

$$\sum_{i=1}^{n}\left(Y_i - \mathbf{W}_i^T\boldsymbol{\gamma} - \boldsymbol{\psi}_{i0}^T\boldsymbol{\eta}_0 - \mathbf{c}_{i1}^T\boldsymbol{\eta}_1 - G_i\mathbf{c}_{i2}^T\boldsymbol{\eta}_2\right)^2 + \lambda_0\boldsymbol{\eta}_0^T\mathbf{S}_0\boldsymbol{\eta}_0 + \lambda_1\boldsymbol{\eta}_1^T\mathbf{S}_1\boldsymbol{\eta}_1 + \lambda_2\boldsymbol{\eta}_2^T\mathbf{S}_2\boldsymbol{\eta}_2.$$

The smoothing parameters now play a key role in model flexibility and the choice of the order and knot selection is no longer that important. The penalized negative log-likelihood can be minimized using least squares and the smoothing parameters are estimated by minimizing the REML criterion (Wood, 2011). This entire estimation can be done using the `mgcv` package in R.

Once the estimates for the $\boldsymbol{\eta}$'s are obtained, we can reconstruct the estimates of the parameter functions using (4.2).

To estimate the standard errors, first note that (4.3) can be rewritten in matrix form as $\mathbf{Y} = \mathbf{C}\boldsymbol{\eta}+\boldsymbol{\varepsilon}$ where $\mathbf{Y} = (Y_1, ..., Y_n)^T$, $\mathbf{C}$ is an $n\times(q+J+K+L)$ matrix with rows $(\mathbf{W}_i^T, \boldsymbol{\psi}_{i0}^T, \mathbf{c}_{i1}^T, \mathbf{c}_{i2}^T)$, $\boldsymbol{\eta} = (\boldsymbol{\gamma}^T, \boldsymbol{\eta}_0^T, \boldsymbol{\eta}_1^T, \boldsymbol{\eta}_2^T)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$. The penalized negative log-likelihood for the Gaussian responses in matrix form is

$$(\mathbf{Y} - \mathbf{C}\boldsymbol{\eta})^T(\mathbf{Y} - \mathbf{C}\boldsymbol{\eta}) + \lambda_0\boldsymbol{\eta}^T\mathbf{D}_0\boldsymbol{\eta} + \lambda_1\boldsymbol{\eta}^T\mathbf{D}_1\boldsymbol{\eta} + \lambda_2\boldsymbol{\eta}^T\mathbf{D}_2\boldsymbol{\eta},$$

where the penalty matrix $\mathbf{D}_m$ is block diagonal containing its corresponding $\mathbf{S}_m$ matrix in the appropriate block and zeros everywhere else. Taking the derivative with respect to $\boldsymbol{\eta}$ yields

$$\hat{\boldsymbol{\eta}} = (\mathbf{C}^T\mathbf{C} + \mathbf{D})^{-1}\mathbf{C}^T\mathbf{y},$$

where $\mathbf{D} = \sum_m \lambda_m\mathbf{D}_m$. Following section 4.8 of Wood (2006), we compute the Bayesian posterior covariance matrix as $\text{var}(\hat{\boldsymbol{\eta}}) = (\mathbf{C}^T\mathbf{C}+\mathbf{D})^{-1}\hat{\sigma}^2$ where $\hat{\sigma}^2$ is estimated from the residual sum of squares. We chose to use the Bayesian approach over the frequentist counterpart because the distributional results can be used directly. We then construct the pointwise standard errors as

$$\begin{aligned}
\text{SE}\{\hat{\beta}_0(g)\} &= \left[\{\boldsymbol{\psi}_0(g)\}^T\text{var}(\hat{\boldsymbol{\eta}}_0)\boldsymbol{\psi}_0(g)\right]^{1/2}, \\
\text{SE}\{\hat{\beta}_1(t)\} &= \left[\{\boldsymbol{\psi}_1(t)\}^T\text{var}(\hat{\boldsymbol{\eta}}_1)\boldsymbol{\psi}_1(t)\right]^{1/2}, \\
\text{SE}\{\hat{\beta}_2(t)\} &= \left[\{\boldsymbol{\psi}_2(t)\}^T\text{var}(\hat{\boldsymbol{\eta}}_2)\boldsymbol{\psi}_2(t)\right]^{1/2},
\end{aligned}$$

where $\boldsymbol{\psi}_0(\cdot) = \{\psi_{01}(\cdot), ..., \psi_{0J}(\cdot)\}^T$, $\boldsymbol{\psi}_1(\cdot) = \{\psi_{11}(\cdot), ..., \psi_{1K}(\cdot)\}^T$, $\boldsymbol{\psi}_2(\cdot) = \{\psi_{21}(\cdot), ..., \psi_{2L}(\cdot)\}^T$, and $\text{var}(\hat{\boldsymbol{\eta}}_m)$ is the appropriate sub-matrix of the variance-covariance matrix of all the parameters $\text{var}(\hat{\boldsymbol{\eta}})$.

4.3 General Functional Model for Interacting Covariates

In practice, the functional linear model can be somewhat restrictive just as the traditional linear model can be in a nonfunctional setting. When the true interaction effect is nonlinear, fitting (4.1) can lead to serious bias, as shown in Section 4.6. A more general model that can handle more complex interactions is

$$Y_i = \mathbf{W}_i^T\boldsymbol{\gamma} + \beta_0(G_i) + \int_{t\in\mathcal{T}} X_i(t)\beta(G_i, t)\,dt + \varepsilon_i, \tag{4.4}$$

where $\beta(g,t)$ is a smooth, unknown parameter function. The linear model is a special case of the one above when $\beta(g,t) = \beta_1(t) + g\beta_2(t)$. In (4.4), the main effect of the scalar covariate G is still modeled nonparametrically by $\beta_0(g)$ and both the main effect of X_i, and the interaction between G_i and X_i are captured in $\beta(g,t)$. This is a more flexible way to model the covariates so that they are not restricted to a linear relationship, and when the true model is linear, using the general model works almost as well as fitting a linear model. Our goal again is to estimate γ, $\beta_0(g)$ and $\beta(g,t)$, and we present two ways of fitting this model.

4.3.1 Estimation by the Empirical Basis

Our first approach is to use a univariate basis expansion for the bivariate function $\beta(g,t)$. We first approximate $X_i(t)$ using FPCA and obtain $\{\hat{\phi}_1(t), ..., \hat{\phi}_P(t)\}$. To motivate a univariate basis expansion of $\beta(g,t)$, recall the typical approach in (4.2) where the unknown function is approximated by the sum of a known set of basis functions multiplied by a set of unknown coefficients. In the case of a bivariate function, we can choose a set of basis functions for the variable t and we can think of the coefficients as fixed realizations of set of unknown functions for G. For example, consider the case where G_i is fixed at g_0 for the ith individual. Now $\beta(g_0,t)$ is a function of only one variable, and we can use the set of orthonormal basis functions obtained from FPCA as the basis expansion for t. Thus the univariate basis expansion is

$$\beta(g_0,t) = b_1(g_0)\hat{\phi}_1(t) + b_2(g_0)\hat{\phi}_2(t) + \cdots + b_P(g_0)\hat{\phi}_P(t),$$

where $\{b_p(g_0)\}_{p=1}^P$ are coefficients depending on g_0, with corresponding eigenfunction basis $\hat{\Phi}(t) = \{\hat{\phi}_p(t)\}_{p=1}^P$. With this idea in mind, when g is arbitrary we write the expansion of β as

$$\beta(g,t) = b_1(g)\hat{\phi}_1(t) + b_2(g)\hat{\phi}_2(t) + \cdots + b_P(g)\hat{\phi}_P(t), \tag{4.5}$$

where the coefficients are functions of g. We are making the crucial assumption that $X(t)$ and $\beta(g,t)$ have the same degree of smoothness in the t direction, which can be restrictive, but the benefits of this estimation procedure are that it simplifies model fitting, there is software readily available for fitting VCMs using spline-based methods (see e.g. `gam` in the `mgcv` package), and as an alternative to spline fitting, one could estimate using local-polynomial kernel smoothing (Müller, 1987; Fan et al., 1998).

Because the set of eigenfunctions form an orthonormal basis, substituting (4.5) into the general model in (4.4) yields a standard varying coefficient model

$$Y_i = \mathbf{W}_i^T\boldsymbol{\gamma} + \beta_0(G_i) + \hat{\xi}_{i1}b_1(G_i) + \hat{\xi}_{i2}b_2(G_i) + \cdots + \hat{\xi}_{iP}b_P(G_i) + \varepsilon_i,$$

where the hats indicate that the ξ's are estimated using FPCA. The penalized negative log-likelihood for this model is

$$\sum_{i=1}^{n}\left\{Y_i - \mathbf{W}_i^T\boldsymbol{\gamma} - \beta_0(G_i) - \sum_{p=1}^{P}\hat{\xi}_{ip}b_p(G_i)\right\}^2 + \lambda_0\int\left\{\beta_0''(g)\right\}^2 dg + \lambda\sum_{p=1}^{P}\int\left\{b_p''(g)\right\}^2 dg,$$

where λ_0 and λ are smoothing parameters. From here, we use a similar technique as in Section 4.2 to fit the model by using B-spline basis approximations $\beta_0(g) = \sum_{j=1}^{J}\eta_{0j}\psi_{0j}(g)$ and $b_p(g) = \sum_{k=1}^{K_p}\eta_{pk}\psi_{pk}(g)$ for $p = 1, ..., P$. When making these substitutions, after some manipulations our the mean structure of the response has the form

$$\begin{aligned} E(Y_i|\mathbf{W}_i, G_i, \hat{X}_i) &= \mathbf{W}_i^T\boldsymbol{\gamma} + \sum_{j=1}^{J}\psi_{0j}(G_i)\eta_{0j} + \int\left\{\hat{\theta}(t) + \sum_{p=1}^{P}\hat{\xi}_{ip}\hat{\phi}_p(t)\right\}\left\{\sum_{p=1}^{P}b_p(G_i)\hat{\phi}_p(t)\right\}dt \\ &= \mathbf{W}_i^T\boldsymbol{\gamma} + \boldsymbol{\psi}_{i0}^T\boldsymbol{\eta}_0 + \sum_{p=1}^{P}(c_p + \hat{\xi}_{ip})\boldsymbol{\psi}_{ip}^T\boldsymbol{\eta}_p, \end{aligned} \tag{4.6}$$

where $c_p = \int\hat{\theta}(t)\hat{\phi}_p(t)\,dt$, $\boldsymbol{\psi}_{ip} = \{\psi_{p1}(G_i), ..., \psi_{pK_p}(G_i)\}^T$ is a $K_p \times 1$ vector, and $\boldsymbol{\eta}_p = (\eta_{p1}, ..., \eta_{pK_p})^T$.

The model in (4.6) is a linear model with $q + J + \sum_p K_p$ parameters. As in the linear model, we can write the penalties in matrix form and thus, the penalized negative log-likelihood is

$$\sum_{i=1}^{n}\left\{Y_i - \mathbf{W}_i^T\boldsymbol{\gamma} - \boldsymbol{\psi}_{i0}^T\boldsymbol{\eta}_0 - \sum_{p=1}^{P}(c_p + \hat{\xi}_{ip})\boldsymbol{\psi}_{ip}^T\boldsymbol{\eta}_p\right\}^2 + \lambda_0\boldsymbol{\eta}_0^T\mathbf{S}_0\boldsymbol{\eta}_0 + \lambda\boldsymbol{\eta}_b^T\mathbf{S}\boldsymbol{\eta}_b,$$

where $\mathbf{S}_0$ is defined as before, and $\mathbf{S}$ is a block diagonal matrix with blocks $\mathbf{S}_p$ where the (i, j)th element of $\mathbf{S}_p$ is $\int\psi_{pi}''(g)\psi_{pj}''(g)\,dg$, and $\boldsymbol{\eta}_b = (\boldsymbol{\eta}_1^T, ..., \boldsymbol{\eta}_P^T)^T$ is a $\sum_p K_p \times 1$ vector where $\boldsymbol{\eta}_p = (\eta_{p1}, ..., \eta_{pK_p})^T$. Thus $\boldsymbol{\eta}_b$ is the parameter vector corresponding to the b_p functions. The same estimating procedure as in the linear case is used, where we minimize the penalized negative log-likelihood and estimate the smoothing parameters by minimizing the REML. We then reconstruct $\beta(g, t)$ as

$$\hat{\beta}(g, t) = \sum_{k=1}^{K_1}\hat{\eta}_{1k}\psi_{1k}(g)\hat{\phi}_1(t) + \sum_{k=1}^{K_2}\hat{\eta}_{2k}\psi_{2k}(g)\hat{\phi}_2(t) + \cdots + \sum_{k=1}^{K_P}\hat{\eta}_{Pk}\psi_{Pk}(g)\hat{\phi}_P(t).$$

The procedure for obtaining the standard errors is similar to that of the linear model. We rewrite (4.6) in matrix form as $\mathbf{Y} = \mathbf{C}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$ but in this case $\mathbf{C}$ has rows $\mathbf{C}_i = \{\mathbf{W}_i^T, \boldsymbol{\psi}_{i0}^T, (c_1 + \hat{\xi}_{i1})\boldsymbol{\psi}_{i1}^T, ..., (c_P + \hat{\xi}_{iP})\boldsymbol{\psi}_{iP}^T\}$ and $\boldsymbol{\eta} = (\boldsymbol{\gamma}^T, \boldsymbol{\eta}_0^T, \boldsymbol{\eta}_b^T)^T$. The penalized negative log-likelihood in

matrix form is

$$(\mathbf{Y} - \mathbf{C}\boldsymbol{\eta})^T(\mathbf{Y} - \mathbf{C}\boldsymbol{\eta}) + \lambda_0 \boldsymbol{\eta}^T \mathbf{D}_0 \boldsymbol{\eta} + \lambda \boldsymbol{\eta}^T \mathbf{S} \boldsymbol{\eta},$$

where $\mathbf{D}_0$ again contains $\mathbf{S}_0$ in the appropriate block and padded with zeros everywhere else. Taking the derivative with respect to $\boldsymbol{\eta}$ yields

$$\hat{\boldsymbol{\eta}} = (\mathbf{C}^T\mathbf{C} + \lambda_0 \mathbf{D}_0 + \lambda \mathbf{S})^{-1} \mathbf{C}^T \mathbf{y}.$$

The Bayesian posterior covariance matrix is $\text{var}(\hat{\boldsymbol{\eta}}) = (\mathbf{C}^T\mathbf{C} + \lambda_0 \mathbf{D}_0 + \lambda \mathbf{S})^{-1} \hat{\sigma}^2$ and the pointwise standard errors for $\hat{\beta}_0(\cdot)$ and $\hat{b}_p(\cdot)$ are

$$\text{SE}\{\hat{\beta}_0(g)\} = \left[\{\boldsymbol{\psi}_0(g)\}^T \text{var}(\hat{\boldsymbol{\eta}}_0) \boldsymbol{\psi}_0(g)\right]^{1/2},$$
$$\text{SE}\{\hat{b}_p(g)\} = \left[\{\boldsymbol{\psi}_p(g)\}^T \text{var}(\hat{\boldsymbol{\eta}}_p) \boldsymbol{\psi}_p(g)\right]^{1/2},$$

where $\boldsymbol{\psi}_p(\cdot) = \{\psi_{p1}(\cdot), ..., \psi_{pK_p}(\cdot)\}^T$ for $p = 1, ..., P$. Define the matrix

$$\mathbf{M}^T = \{\boldsymbol{\phi}(t)\}^T \text{diag}[\{\boldsymbol{\psi}_1(g)\}^T, ..., \{\boldsymbol{\psi}_P(g)\}^T],$$

where $\boldsymbol{\phi}(t) = \{\phi_1(t), \phi_2(t), ..., \phi_P(t)\}^T$. Then $\hat{\beta}(g,t) = \mathbf{M}^T \hat{\boldsymbol{\eta}}_b$ and the pointwise standard error is $\text{SE}\{\hat{\beta}(g,t)\} = \mathbf{M}^T \text{var}(\hat{\boldsymbol{\eta}}_b) \mathbf{M}$.

4.3.2 Estimation using a Tensor Product Basis

In practice, we do not know if the functional covariate $X(t)$ and the parameter function $\beta(g,t)$ have the same smoothness, so an alternative to the empirical basis is to use a tensor product basis to model $\beta(g,t)$. Again, we start with FPCA to obtain the estimate $\hat{X}_i(t)$ and write the penalized negative log-likelihood as

$$\sum_{i=1}^{n} \ell\left(Y_i; \mathbf{W}_i, G_i, \hat{X}_i\right) + \lambda_0 \int \{\beta_0''(g)\}^2 \, dg + \lambda_1 \int \left\{\frac{\partial^2}{\partial g^2} \beta(g,t)\right\}^2 dg \, dt + \lambda_2 \int \left\{\frac{\partial^2}{\partial t^2} \beta''(g,t)\right\}^2 dg \, dt$$

where $\ell(Y_i; \mathbf{W}_i, G_i, \hat{X}_i) = \{Y_i - \mathbf{W}_i^T \boldsymbol{\gamma} - \beta_0(G_i) - \int \hat{X}_i(t) \beta(G_i, t) \, dt\}^2$. For the bivariate function, we smooth in both the g and t directions and use two penalty parameters, λ_1 and λ_2, to allow each direction to have a different amount of smoothness. We use the basis representation $\beta_0(g) = \sum_{j=1}^{J} \eta_{0j} \psi_{0j}(g)$ as before, and for $\beta(g,t)$, we use a tensor product between two B-splines,

$$\beta(g,t) = \sum_{k=1}^{K} \sum_{l=1}^{L} \psi_{1k}(g) \psi_{2l}(t) \eta_{kl}. \tag{4.7}$$

Here we use two different sets of B-splines for g and t with cross products indexed by an unknown parameter η_{kl}. Using these expansions, (4.4) can be written as

$$Y_i = \mathbf{W}_i^T \boldsymbol{\gamma} + \sum_{j=1}^{J} \psi_{0j}(G_i)\eta_{0j} + \sum_{k=1}^{K}\sum_{l=1}^{L} c_{ikl}\eta_{kl} + \varepsilon_i, \tag{4.8}$$

where $c_{ikl} = \psi_{1k}(G_i)\int \hat{X}_i(t)\psi_{2l}(t)\,dt$ and $\boldsymbol{\gamma}$, η_{0j}, and η_{kl} are unknown parameters. Once again we are reduced to a linear model with $q + J + KL$ parameters.

By substituting in the basis approximations and writing the penalty terms in matrix form as before, the penalized negative log-likelihood has the form

$$\sum_{i=1}^{n} \ell(Y_i; \mathbf{W}_i, G_i, \hat{X}_i) + \lambda_0 \boldsymbol{\eta}_0^T \mathbf{S}_0 \boldsymbol{\eta}_0 + \lambda_1 \boldsymbol{\eta}_{\text{KL}}^T (\mathbf{S}_1 \otimes \mathbf{I}_L)\boldsymbol{\eta}_{\text{KL}} + \lambda_2 \boldsymbol{\eta}_{\text{KL}}^T (\mathbf{I}_K \otimes \mathbf{S}_2)\boldsymbol{\eta}_{\text{KL}},$$

where $\boldsymbol{\eta}_{\text{KL}}$ is a $KL \times 1$ vector that is grouped along the K index. That is, $\boldsymbol{\eta}_{\text{KL}} = (\boldsymbol{\eta}_1^T, ..., \boldsymbol{\eta}_K^T)^T$ where $\boldsymbol{\eta}_k = (\eta_{k1}, ..., \eta_{kL})^T$ for $k = 1, ..., K$ (See Wood (2006) Section 4.1.8). The (i, j)th element of $\mathbf{S}_1$ is $\int \psi''_{1i}(g)\psi''_{1j}(g)\,dg$ and similarly, the (i, j)th element of $\mathbf{S}_2$ is $\int \psi''_{2i}(t)\psi''_{2j}(t)\,dt$. We minimize the penalized negative log-likelihood and estimate the smoothing parameters using REML as before, then reconstruct $\beta(g, t)$ as

$$\hat{\beta}(g, t) = \sum_{k=1}^{K}\sum_{l=1}^{L} \psi_{1k}(g)\psi_{2l}(t)\hat{\eta}_{kl}.$$

To compute the standard errors, observe that the penalized negative log-likelihood has the form

$$(\mathbf{Y} - \mathbf{C}\boldsymbol{\eta})^T(\mathbf{Y} - \mathbf{C}\boldsymbol{\eta}) + \lambda_0 \boldsymbol{\eta}^T \mathbf{D}_0 \boldsymbol{\eta} + \lambda_1 \boldsymbol{\eta}^T \mathbf{D}_1 \boldsymbol{\eta} + \lambda_2 \boldsymbol{\eta}^T \mathbf{D}_2 \boldsymbol{\eta},$$

where $\mathbf{C}$ has rows $(\mathbf{W}_i^T, \boldsymbol{\psi}_{0i}^T, \mathbf{c}_{i\text{KL}})$, where $\mathbf{c}_{i\text{KL}}$ is grouped in the same way as $\boldsymbol{\eta}_{\text{KL}}$, $\mathbf{D}_1 = \mathbf{S}_1 \otimes \mathbf{I}_L$, $\mathbf{D}_2 = \mathbf{I}_K \otimes \mathbf{S}_2$, and $\boldsymbol{\eta} = (\boldsymbol{\gamma}^T, \boldsymbol{\eta}_0^T, \boldsymbol{\eta}_{\text{KL}}^T)^T$. Exactly as before, taking the derivative results in $\hat{\boldsymbol{\eta}} = (\mathbf{C}^T\mathbf{C} + \mathbf{D})^{-1}\mathbf{C}^T\mathbf{y}$ where $\mathbf{D} = \sum_i \lambda_i \mathbf{D}_i$ and the Bayesian posterior covariance matrix is $\text{var}(\hat{\boldsymbol{\eta}}) = (\mathbf{C}^T\mathbf{C} + \mathbf{D})^{-1}\hat{\sigma}^2$. The pointwise standard error of $\hat{\beta}(g, t)$ is constructed as

$$\text{SE}\{\hat{\beta}(g, t)\} = \left[\{\boldsymbol{\psi}_1(g)\}^T \otimes \{\boldsymbol{\psi}_2(t)\}^T\right]^T \text{var}(\hat{\boldsymbol{\eta}}_{\text{KL}}) \left[\{\boldsymbol{\psi}_1(g)\}^T \otimes \{\boldsymbol{\psi}_2(t)\}^T\right],$$

where again, $\text{var}(\hat{\boldsymbol{\eta}}_{\text{KL}})$ is the appropriate sub-matrix of the covariance of the entire parameter vector $\boldsymbol{\eta}$.

4.4 Testing

Testing in this framework is difficult due to model complexity and the added complications from estimating the smoothing parameters. Thus we present a naive test in this chapter and stress to the reader that this is still a very open problem that requires further investigation than is dedicated here. In order to test for the effects of the covariates and for the interaction between G and X, we follow the procedures from Section 4.8.5 of Wood (2006) and use the Wald statistic to establish a test in the linear model framework, and in the general model framework using the tensor product basis fit. Testing in the empirical basis framework is more complicated and still under investigation.

Recall that, after some manipulations, the models in (4.3) and (4.8) can be written in a general form as $\mathbf{Y} = \mathbf{C}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, where $\mathbf{C}$ is the appropriate design matrix and $\boldsymbol{\eta}$ is a vector containing all of the parameters. Note that the structures of $\mathbf{C}$ and $\boldsymbol{\eta}$ are different for the two models, but the testing procedure is similar, once they are written in this form. Testing for certain effects in the original models is equivalent to testing if some subset $\boldsymbol{\eta}_m$ of $\boldsymbol{\eta}$ is identically zero. Under the null, we have that $\hat{\boldsymbol{\eta}}_m \sim \mathrm{N}(\mathbf{0}, \mathbf{V}_{\hat{\boldsymbol{\eta}}_m})$ where $\mathbf{V}_{\hat{\boldsymbol{\eta}}_m}$ is the appropriate sub-matrix of $\mathrm{var}(\hat{\boldsymbol{\eta}})$, and if $\mathbf{V}_{\hat{\boldsymbol{\eta}}_m}$ is full rank, then it follows that

$$\hat{\boldsymbol{\eta}}_m^T \mathbf{V}_{\hat{\boldsymbol{\eta}}_m}^{-1} \hat{\boldsymbol{\eta}}_m \sim \chi^2_{\dim(\hat{\boldsymbol{\eta}}_m)}. \tag{4.9}$$

Even if $\mathbf{V}_{\hat{\boldsymbol{\eta}}_m}$ is not full rank, we can still construct a test statistic based on the pseudoinverse that follows a χ^2 distribution. See Wood (2006) for further details.

When there are no smoothing parameters in the model, or if the smoothing parameters are known, then the test statistic in (4.9) performs reasonably. However, when the smoothing parameters need to be estimates, as is often the case in real world applications, then this test may have lower p-values than they should be. Wood (2006) notes that if the test definitively rejects or fails to reject, then any inference based on the test is most likely reliable. However if the p-values are near the decision threshold, then care needs to be taken when drawing conclusions. In our simulations (see Section 4.6.3) we found the Wald test to be extremely conservative.

Another issue that arises from this extremely naive test statistic is that the number of basis functions chosen plays a significant role. The degrees of freedom of the χ^2 distribution depend on the dimension of the subset of parameters being tested, thus choosing a large number of basis functions can cause the variance of the distribution to increase rapidly resulting in loss of power, especially for the tensor product fit, since the number of basis functions has a multiplicative relationship. A method for appropriately choosing the number of basis functions or developing a better procedure is part of this open problem of testing. Despite the issues discussed, we

present the testing methodology anyway and allow readers to take from it what they will.

To test for effects in the linear model is straightforward. Testing for the interaction between X and G is equivalent to testing $H_0\colon \beta_2(t) = 0$. Recall that a basis expansion is used in (4.2) so that this null is equivalent to $H_0\colon \boldsymbol{\eta}_2 = \mathbf{0}$. We construct the test statistic $\hat{\boldsymbol{\eta}}_2^T \left\{\mathrm{var}(\hat{\boldsymbol{\eta}}_2)\right\}^{-1} \hat{\boldsymbol{\eta}}_2$ and compare it to a χ^2_L to obtain the p-values. If the test for interaction is not significant, then we can investigate the main effects of G and X. This is done by first fitting the main effects model

$$Y_i = \mathbf{W}_i^T \boldsymbol{\gamma} + \beta_0(G_i) + \int X_i(t)\beta_{\mathrm{M}}(t)\, dt + \varepsilon_i, \tag{4.10}$$

where $\beta_{\mathrm{M}}(t)$ is estimated in the same way as $\beta_1(t)$ in Section 4.2. The Wald statistic for testing the main effects of G and X is $\hat{\boldsymbol{\eta}}_0^T \left\{\mathrm{var}(\hat{\boldsymbol{\eta}}_0)\right\}^{-1} \hat{\boldsymbol{\eta}}_0$ and $\hat{\boldsymbol{\eta}}_{\mathrm{M}}^T \left\{\mathrm{var}(\hat{\boldsymbol{\eta}}_{\mathrm{M}})\right\}^{-1} \hat{\boldsymbol{\eta}}_{\mathrm{M}}$, respectively, and are compared to a χ^2_J and $\chi^2_{\dim(\hat{\boldsymbol{\eta}}_{\mathrm{M}})}$, respectively.

Testing for the interaction between G and X in the tensor product model is a little more difficult. Recall that $\beta(g,t)$ contains both the main effect of X and the interaction effect of X and G, so testing for interaction is equivalent to testing that $\beta(g,t)$ does not depend on G, that is, $H_0\colon \beta(g,t) = \beta(t)$. There is no direct subset of $\boldsymbol{\eta}$ that corresponds to this null hypothesis, so we need to do some slight manipulations to the model. Theoretically, we can separate the effects of the bivariate function into a purely main effect term and a purely interaction effect term. Thus let $\beta(g,t) = \beta_{\mathrm{M}}(t) + \beta_{\mathrm{I}}(g,t)$ where $\beta_{\mathrm{M}}(t)$ models the main effect of X and $\beta_{\mathrm{I}}(g,t)$ models the interaction effect of X and G. The basis expansion for the separated function is

$$\begin{aligned}\beta_{\mathrm{M}}(t) + \beta_{\mathrm{I}}(g,t) &\approx \{1, \psi^*_{11}(g), ..., \psi^*_{1K}(g)\} \otimes \{\psi_{21}(t), ..., \psi_{2L}(t)\} \left(\boldsymbol{\eta}_{\mathrm{M}}^T, \boldsymbol{\eta}_{\mathrm{I}}^T\right)^T \\ &= \left[\{1, \boldsymbol{\psi}^*_1(g)\} \otimes \{\boldsymbol{\psi}_2(t)\}\right]^T \left(\boldsymbol{\eta}_{\mathrm{M}}^T, \boldsymbol{\eta}_{\mathrm{I}}^T\right)^T,\end{aligned}$$

where each $\psi^*_{1k}(g)$ is orthogonal to 1, to avoid fitting the main effect twice. Substituting this basis expansion into the general model results in a linear model of the form

$$Y_i = \mathbf{W}_i^T \boldsymbol{\gamma} + \boldsymbol{\psi}_{0i}^T \boldsymbol{\eta}_0 + \mathbf{c}_{i\mathrm{M}}^T \boldsymbol{\eta}_{\mathrm{M}} + \mathbf{c}_{i\mathrm{I}}^T \boldsymbol{\eta}_{\mathrm{I}} + \varepsilon_i,$$

where $\mathbf{c}_{i\mathrm{M}}$ is an $L \times 1$ vector with elements $\int \hat{X}_i(t)\psi_{2l}(t)\, dt$ and $\mathbf{c}_{i\mathrm{I}}$ is a $KL \times 1$ vector with elements $\psi_{1k}(G_i) \int \hat{X}_i(t)\psi_{2l}(t)\, dt$. Now we can test for the interaction effect by testing if $\boldsymbol{\eta}_{\mathrm{I}}$ is identically zero using the Wald statistic as before. If the interaction is not found to be significant, then we can test for the main effects by fitting (4.10).

4.5 Extensions and Generalizations

4.5.1 Exponential Family Responses

For ease of presentation, we assumed that the response was Gaussian, but this does not have to be the case for our estimation methods to work. If the response's conditional distribution belongs to a single-parameter exponential family, the general model in (4.4) can be written as

$$g(\mu_i) = \mathbf{W}_i^T \boldsymbol{\gamma} + \beta_0(G_i) + \int_{t \in \mathcal{T}} X_i(t)\beta(G_i, t)\, dt, \tag{4.11}$$

where $\mu_i = \mathrm{E}(Y_i|\mathbf{W}_i, G_i, X_i)$ is the conditional mean of the response Y_i and $g(\cdot)$ is a link function from the generalized linear model framework from Nelder and Wedderburn (1972). The manipulations in (4.3), (4.6), and (4.8) still apply with the left hand side replaced by $g(\mu_i)$ and the error term omitted. The respective penalized negative log-likelihoods are minimized using penalized iterative reweighted least squares (see e.g. Wood (2006) Section 3.4) and the smoothing parameters are estimated by minimizing the approximate REML criterion (Wood, 2011).

4.5.2 Sparse Observations and Measurement Error

In most real-world situations, we only observe noisy, discrete realizations of the functional covariates and often individuals in the study skip appointments or drop out early resulting in a potential large amount of missing values. This may cause issues in local estimation techniques if some partitions are not dense enough to apply FPCA, however since our estimation methods are global, they are likely not affected by extremely sparse observations. As long as the entire domain of t is dense across all subjects, then even noisy curves with only two or three realizations can still be smoothed and estimated with FPCA.

4.6 Simulation Study

4.6.1 Simulation Setup

We conducted a simulation study to evaluate our proposed methods. We generated three parametric covariates $\mathbf{W}_i = (W_{i1}, W_{i2}, W_{i3})^T$ for each subject from a standard normal distribution and let $\boldsymbol{\gamma} = (1, 0.5, -1)^T$. We generated the interacting covariate G_i uniformly over $[-1, 1]$. We generated the functional covariates $X_i(t)$ using four Fourier basis functions, where the coefficients were normally distributed with mean 0 and variances $(8, 4, 2, 1)$. The random error for the continuous response ε_i was generated from a standard normal distribution. We defined

$\beta_0(g) = (g - 1/2)^2 + 1$ and the form of $\beta(g,t)$ was either from the linear model, the empirical basis (EB) model, or the tensor product (TP) model. For the linear case, $\beta_1(t) = 1$, $\beta_2(t) = 2\sin(\pi t)$, and $\beta(g,t) = \beta_1(t) + g\beta_2(t)$ which coincides with (4.1). For the EB case in (4.5), $b_1(g) = g + 1$, $b_2(g) = g^2$ and $\beta(g,t) = b_1(g)\sin(\pi t) + b_2(g)\cos(\pi t)$. For the TP case in (4.7), $\beta(g,t) = e^{(g-0.2)^2}\{1 + \sin(\pi t)\}/4$. For relativity purposes, the scaled norms were calculated where $\|\beta_0(g)\|_2/2 = 2.93$, and $\|\beta(g,t)\|_2/2$ were 1.67, 0.77, and 0.52 for the linear, EB, and TP surfaces, respectively. Finally, the response Y_i was generated from either (4.4) or (4.11) using a high dimensional grid for t.

For our analysis, we used a noisy functional covariate $Z_i(t) = X_i(t) + \delta_i(t)$ where $\delta_i(t)$ is a white noise process with mean 0 and variance 1/4, and $Z_i(t)$ was observed either on a dense grid of 41 equally spaced timepoints or on a sparse grid of 101 timepoints in [0,1]. For the sparse grid, we randomly selected a number m_i between 16 and 30 for each subject, and then randomly chose m_i of the 101 grid points to be the observed values, setting the rest to missing values. We fit the linear model, EB model, and TP model for all three choices of $\beta(g,t)$ using $J = K = L = 9$ basis functions and compared the results. We expected optimal performance when the true $\beta(g,t)$ is the same as the fitted model.

We estimated the functions on an equally spaced grid for $g \in [-1,1]$ and $t \in [0,1]$. We evaluated our estimates for each function based on the integrated squared bias, mean integrated squared error, empirical variance, estimated variance, and integrated coverage. We defined $\hat{\beta}_0(g) = R^{-1}\sum_r \hat{\beta}_0^{(r)}(g)$ and

$$
\begin{aligned}
\mathrm{SB} &= \frac{1}{2}\int \{\hat{\beta}_0(g) - \beta_0(g)\}^2\, dg, \\
\mathrm{MSE} &= \frac{1}{R}\sum_r \left[\frac{1}{2}\int \{\hat{\beta}_0^{(r)}(g) - \beta_0(g)\}^2\, dg\right], \\
\mathrm{Vem} &= \frac{1}{2}\int \frac{1}{R}\sum_r \{\hat{\beta}_0^{(r)}(g) - \hat{\beta}_0(g)\}^2\, dg, \\
\mathrm{Ves} &= \frac{1}{2}\int \frac{1}{R}\sum_r [\mathrm{SE}\{\hat{\beta}_0^{(r)}(g)\}]^2\, dg, \\
\mathrm{C} &= \frac{1}{2}\int R^{-1}\sum_r I[\hat{\beta}_0(g) \in \mathrm{CI}\{\hat{\beta}_0^{(r)}(g)\}]\, dg,
\end{aligned}
$$

where R is the number of simulation replications and $\mathrm{CI}(x) = (x - 1.96\,\mathrm{SE}(x), x + 1.96\,\mathrm{SE}(x))$ is the 95% confidence interval. All integrals are scaled by the volume of G. These definitions are similar for $\beta(g,t)$ except that we integrate over both g and t and scale by the respective volumes.

4.6.2 Estimation

Continuous Response

The results for the continuous response are presented it Table 4.1. For our estimate of $\hat{\beta}_0(g)$, we did not see much difference in bias across the three different fits for each surface and as sample size increased, the bias decreased. When fitting the linear model, the variability of our estimate for $\hat{\beta}_0(g)$ increased as the structure of $\hat{\beta}(g,t)$ moved from linear to EB to TP. For the other two fits there was no real change in variability across the three surface types, and both fits yielded similar estimated and empirical variance. Thus overall, the EB and TP fits had very similar biases and MISEs. In comparison, the linear fit had lower MISE when the linear surface was used, similar MISE when the EB surface was used, and higher MISE when the TP surface was used. As sample size increased, the integrated coverage for all three fits was similar and close to the nominal 95% value. We saw similar trends for $\hat{\beta}_0(g)$ when the sparse grid of timepoints was used.

For the estimates of $\hat{\beta}(g,t)$ we found that the integrated squared bias for the EB fit was uniformly lower than the other two fits for all surfaces and sample sizes, except for the dense case where the true surface was linear and the sample size was 400. All three fits had extremely low relative bias when the true surface was linear, the highest value being 0.016, corresponding to the TP fit when $n = 100$. When the true surface was nonlinear, the linear fit had uniformly higher relative bias and was especially high when the true surface was from the TP model (e.g. $0.067/0.52 = 12.8\%$). The other two fits performed well in terms of relative bias, the highest being 6% corresponding to the TP fit of the EB. surface when $n = 100$.

In terms of MISE, the linear fit perform the best when the true surface was linear, and the TP fit performed the best for the other two surfaces. The EB fit had high MISE for all three surfaces, due to the uniformly higher amount of variability of the fit. The linear and TP fits had comparable variability with the linear fit doing slightly better for the linear and EB surfaces. The linear fit generally had the worse coverage, even when the true surface was linear, while the EB and TP fits had comparable coverage. For the sparse grid of timepoints, we found a similar trend where the EB fit had uniformly lower bias, and the TP fit had lower MISE and variance when the true surface was not linear.

In summary, we found that the EB fit had uniformly lower bias than the other two fits, but also had uniformly higher variance, causing the MISE to be high. The linear fit performed well when the true surface was linear, but quickly fell apart for the nonlinear surfaces. The TP fit was stable and consistent for all three surfaces, had appropriate coverage, and naturally performed the best when the true surface was TP. Because of this, we feel the TP surface should be used in practice since the true structure of the bivariate surface is often unknown.

Table 4.1: Gaussian response: Integrated squared bias (SB), mean integrated squared error (MSE), integrated Monte Carlo empirical variance (Vem), integrated estimated variance (Ves), and integrated coverage (C) for the two functions $b_0(g)$ and $\beta(g,t)$ when the functional covariate is densely observed and sparsely observed. We fit a linear model (L), an empirical basis model (EB), and a tensor product basis model (TP) to each of the three true surfaces for sample sizes 100, 200, and 400. We also estimated the residual error for the dense (D) and sparse (S) case, which should be close to 1. All values in the table are multiplied by 100. For relativity, the scaled norm for $\beta_0(g)$ is 2.93, and the scaled norms for $\beta(g,t)$ are 1.67, 0.77, and 0.52 for the linear, EB, and TP surfaces, respectively.

		$\beta_0(g)$ Dense					$\beta(g,t)$ Dense					$\beta_0(g)$ Sparse					$\beta(g,t)$ Sparse					Residual	
		SB	MSE	Vem	Ves	C	SB	MSE	Vem	Ves	C	SB	MSE	Vem	Ves	C	SB	MSE	Vem	Ves	C	D	S
LINEAR SURFACE																							
L	100	0.4	4.8	4.4	4.1	93.4	1.1	10.9	9.8	8.6	89.7	0.4	5.3	4.9	4.3	92.4	0.8	15.6	14.8	13.1	88.6	87.8	91.6
	200	0.1	2.2	2.1	2.2	95.5	0.4	5.5	5.1	5.6	93.4	0.1	2.3	2.2	2.3	95.4	0.2	6.5	6.3	6.6	93.3	94.2	96.5
	400	0.1	1.2	1.1	1.2	96.1	0.1	2.6	2.5	4.0	97.3	0.1	1.2	1.1	1.2	96.1	0.1	3.2	3.1	4.2	96.6	97.4	99.3
EB	100	0.5	5.2	4.8	4.1	92.4	0.3	22.1	21.8	18.1	93.8	0.5	5.1	4.6	4.3	93.8	0.4	106.4	106.1	66.0	92.2	85.9	86.7
	200	0.2	2.4	2.2	2.2	95.1	0.2	9.4	9.2	8.3	94.8	0.1	2.4	2.3	2.3	94.9	0.2	38.7	38.6	26.4	93.0	92.8	95.0
	400	0.1	1.2	1.1	1.2	95.7	0.3	4.8	4.6	4.1	94.7	0.1	1.2	1.1	1.2	96.5	0.1	14.5	14.4	10.2	93.2	97.2	98.5
TP	100	0.4	5.1	4.7	4.0	92.5	2.7	12.0	9.2	8.1	90.2	0.4	5.1	4.7	4.2	93.1	2.1	15.4	13.3	13.3	88.6	86.8	90.6
	200	0.1	2.3	2.2	2.2	95.1	1.2	6.7	5.6	6.2	93.3	0.1	2.4	2.3	2.3	94.8	1.0	7.6	6.7	7.0	93.2	93.5	95.6
	400	0.1	1.2	1.1	1.2	96.2	0.3	3.3	3.0	4.9	97.7	0.1	1.2	1.1	1.2	95.7	0.3	3.9	3.6	5.1	97.3	97.3	98.9
EMPIRICAL BASIS SURFACE																							
L	100	0.4	5.3	5.0	4.3	92.6	6.8	18.3	11.4	9.0	87.1	0.4	5.3	4.9	4.3	92.3	6.5	21.2	14.8	12.0	86.1	93.0	92.9
	200	0.2	2.5	2.4	2.3	95.1	5.9	12.4	6.5	6.1	91.4	0.2	2.6	2.4	2.3	95.3	5.8	13.7	7.8	6.8	89.5	99.8	99.9
	400	0.1	1.2	1.2	1.3	96.1	5.4	8.8	3.4	4.3	95.8	0.1	1.2	1.1	1.3	96.3	5.4	9.8	4.4	4.6	94.3	102.6	103.4
EB	100	0.5	5.3	4.8	4.2	92.6	2.1	26.3	24.2	20.0	93.2	0.4	5.6	5.2	4.3	92.4	2.5	103.3	100.9	67.0	92.9	85.1	84.1
	200	0.2	2.4	2.2	2.3	95.1	1.0	12.4	11.4	10.4	94.0	0.1	2.4	2.2	2.3	95.4	1.0	40.7	39.8	29.8	93.5	91.5	91.9
	400	0.1	1.2	1.1	1.2	96.0	0.5	6.0	5.5	5.4	95.4	0.1	1.2	1.1	1.2	96.2	0.4	16.4	16.0	12.6	94.3	95.7	96.0
TP	100	0.5	5.4	4.9	4.1	92.2	4.5	15.6	11.1	9.0	87.7	0.5	5.2	4.7	4.1	92.8	3.8	18.0	14.3	11.2	87.1	86.9	87.8
	200	0.1	2.4	2.3	2.2	94.9	1.8	8.3	6.5	7.7	93.6	0.1	2.5	2.3	2.3	94.8	1.7	9.2	7.5	8.2	92.0	92.5	93.0
	400	0.0	1.2	1.1	1.2	96.1	0.8	4.2	3.4	6.2	97.9	0.1	1.2	1.1	1.2	96.2	0.7	4.6	3.9	6.4	97.5	95.8	96.0
TENSOR PRODUCT SURFACE																							
L	100	0.4	8.9	8.4	5.7	91.3	6.8	19.7	13.0	8.2	89.7	0.5	9.5	9.0	5.7	90.4	6.9	27.4	20.5	16.1	90.3	129.6	130.3
	200	0.3	4.4	4.1	3.1	92.6	6.8	13.2	6.4	4.3	89.0	0.2	4.3	4.1	3.1	92.8	6.8	14.0	7.2	4.6	88.8	140.6	140.9
	400	0.1	2.2	2.1	1.7	94.5	6.7	9.9	3.2	2.2	87.5	0.1	2.4	2.3	1.7	93.5	6.7	10.2	3.6	2.4	86.4	146.7	146.5
EB	100	0.5	5.4	4.9	4.3	92.6	0.3	22.9	22.6	23.7	96.0	0.4	5.3	4.9	4.4	92.9	0.4	101.7	101.4	70.8	95.4	81.3	81.2
	200	0.2	2.5	2.4	2.3	94.7	0.2	11.7	11.5	11.8	96.5	0.2	2.3	2.2	2.3	95.9	0.2	40.9	40.7	30.7	95.6	90.0	89.5
	400	0.1	1.2	1.1	1.2	95.7	0.1	6.1	6.0	6.5	96.7	0.1	1.2	1.1	1.2	96.4	0.1	16.6	16.5	13.9	96.4	94.6	95.0
TP	100	0.4	5.3	4.9	4.2	92.9	1.0	7.5	6.5	6.5	96.3	0.5	5.3	4.9	4.2	92.8	0.9	9.2	8.3	7.8	96.1	84.1	84.4
	200	0.2	2.3	2.2	2.2	95.5	0.8	4.5	3.7	3.9	95.8	0.2	2.4	2.2	2.2	95.5	0.8	4.6	3.8	3.9	95.6	91.4	91.6
	400	0.1	1.2	1.1	1.2	96.2	0.6	2.7	2.1	2.4	94.4	0.1	1.1	1.1	1.2	96.4	0.6	2.9	2.3	2.5	94.5	95.2	96.4

Binary Response

When the response was binary, the results from our simulation (Table 4.2) were more unstable due to sample size issues, thus we used relatively large sample sizes of 300 and 500. When the true surface was linear, we found that the linear and TP estimate of $\beta_0(g)$ performed reasonably well for the dense and sparse cases and had similar results. The EB fit generally had the highest bias and variability in the estimate of $\beta_0(g)$ and had the worst coverage. For the estimate of the linear $\beta(g,t)$, we found that the TP fit generally had the highest bias and the lowest variability and MSE, but was comparable to the linear fit overall. The EB surface had the lowest bias for large sample sizes, but had three times as much variability in the dense case, and six times as much variability in the sparse case, compared to the other two fits.

When the true surface had the EB structure, we again saw very comparable results between the linear and TP estimation of $\beta_0(g)$. The EB fit had uniformly higher variability and tended to have the worst coverage of all three fits. For the estimate of $\beta(g,t)$, we saw slightly elevated bias in the linear fit compared to the other two fits, with overall relative bias around 10%. In comparison, the relative bias for the EB and TP fits were around 4% and 7%, respectively. In terms of variation, the linear and TP fits had similar results, with the TP fit doing slightly better. The EB fit had uniformly higher variability which resulted in better integrated coverage than the other two fits.

When the true surface had the TP type, we found a noticeable increase in the bias of $\beta_0(g)$ when the linear estimation was used. This is because the excess nonlinear effects that are not captured by the linear fit is absorbed into the main effects term, which results in added bias. Also of note is that, as sample size increased, the bias of the linear estimate of $\beta_0(g)$ also increased, showing that this bias is not a sample size issue. However, even though the linear fit had as much as 15 times the relative bias as the other two fits, the overall relative bias was only 3%. We found that the linear estimate of $\beta_0(g)$ had uniformly lower variance than the other two fits while the TP fit had uniformly lower MISE. The EB estimate had the lowest relative bias (less that 1%) and the highest variability.

When we considered our estimate of $\beta(g,t)$ for TP surface, we found that the linear fit naturally had high relative bias compared to the other two estimates with relative bias as high as 15%, corresponding to $n = 500$ using the dense grid. In comparison, the relative bias of the TP estimate was 4%, while the EB fit had the lowest relative bias at 2% of the scaled norm. Again, we found that the EB fit had the uniformly highest variability and MISE for both the dense and sparse designs which overshadowed any gains in bias when using this method. The TP fit had uniformly lower MISE and similar variance to the linear estimation, with the linear

Table 4.2: Binary response: Integrated squared bias (SB), mean integrated squared error (MSE), integrated MC empirical variance (Vem), integrated estimated variance (Ves), and integrated coverage (C) for the two functions $b_0(g)$ and $\beta(g,t)$ when the functional covariate is densely observed and sparsely observed. We fit a linear model (L), an empirical basis model (EB), and a tensor product basis model (TP) to each of the three true surfaces for sample sizes 300, and 500. All values in the table are multiplied by 100. For relativity, the scaled norm for $\beta_0(g)$ is 2.93, and the scaled norms for $\beta(g,t)$ are 1.67, 0.77, and 0.52 for the linear, EB, and TP surfaces, respectively.

		$\beta_0(g)$ Dense					$\beta(g,t)$ Dense					$\beta_0(g)$ Sparse					$\beta(g,t)$ Sparse				
		SB	MSE	Vem	Ves	C	SB	MSE	Vem	Ves	C	SB	MSE	Vem	Ves	C	SB	MSE	Vem	Ves	C
LINEAR SURFACE																					
L	300	2.3	26.9	24.6	19.9	93.0	3.4	45.1	41.8	32.0	91.7	3.3	30.4	27.2	20.6	92.4	2.8	56.2	53.4	38.5	91.5
	500	0.9	14.3	13.4	11.4	93.2	2.0	26.5	24.5	19.1	90.8	1.0	14.3	13.3	11.3	93.5	2.6	26.6	24.1	18.9	91.1
EB	300	4.1	33.6	29.5	21.5	92.5	4.0	122.7	118.9	96.5	94.6	5.2	37.5	32.3	22.5	91.4	5.3	474.6	469.8	265.3	93.9
	500	1.3	16.0	14.7	11.7	92.4	1.2	59.6	58.4	50.2	94.8	1.5	17.0	15.6	12.0	92.6	1.4	177.4	176.1	116.6	94.1
TP	300	2.8	30.0	27.2	19.4	91.9	6.2	44.5	38.3	29.6	93.2	2.8	30.1	27.3	19.4	92.3	5.5	49.9	44.4	31.9	92.9
	500	1.2	14.3	13.1	11.0	93.3	4.7	25.7	21.0	17.2	92.8	0.9	13.9	13.0	11.1	94.0	5.1	28.8	23.7	17.3	92.2
EMPIRICAL BASIS SURFACE																					
L	300	1.8	21.2	19.4	15.8	92.8	7.3	42.5	35.2	25.5	90.0	1.8	21.4	19.7	15.7	92.6	8.1	45.8	37.7	26.5	90.6
	500	0.7	11.0	10.4	9.1	93.3	7.4	25.5	18.1	14.3	89.7	0.8	11.0	10.2	8.9	93.8	7.1	27.3	20.2	15.3	89.9
EB	300	2.5	26.3	23.8	16.6	92.1	3.5	91.6	88.2	74.0	94.5	3.3	29.3	26.0	17.4	92.1	3.1	320.7	318.0	196.0	94.1
	500	0.7	12.7	12.0	9.6	92.6	2.5	46.8	44.4	40.2	95.2	0.9	12.3	11.4	9.7	93.3	2.4	140.6	138.3	94.8	94.4
TP	300	2.0	22.1	20.1	15.4	92.6	5.6	36.2	30.6	23.4	91.2	1.5	20.6	19.1	15.3	93.3	5.9	37.5	31.7	24.0	91.9
	500	0.8	11.2	10.4	9.0	93.6	5.3	22.6	17.2	14.4	90.9	0.7	12.0	11.3	9.1	92.3	5.2	24.6	19.4	15.2	90.9
TENSOR PRODUCT SURFACE																					
L	300	10.0	21.5	11.5	10.8	94.4	7.7	26.3	18.6	14.3	94.6	10.3	22.3	12.0	11.1	94.0	7.5	26.8	19.3	15.1	94.2
	500	11.4	18.3	6.9	6.2	94.2	7.9	18.3	10.4	8.1	94.1	10.9	17.8	6.9	6.2	93.6	7.8	19.6	11.8	8.9	94.5
EB	300	2.5	23.0	20.6	15.3	92.9	1.0	79.8	78.8	69.6	95.6	2.6	23.7	21.1	16.2	92.7	1.4	273.5	272.5	179.9	95.3
	500	1.8	12.5	10.7	8.9	93.5	1.0	39.2	38.3	37.5	96.1	1.6	12.6	11.0	9.1	92.7	1.0	114.1	113.2	85.7	95.6
TP	300	2.4	20.4	18.1	14.1	93.5	2.0	20.5	18.5	17.1	95.7	2.8	20.0	17.2	13.9	93.1	2.0	23.4	21.4	18.1	95.9
	500	1.9	12.2	10.2	8.6	93.1	1.6	13.0	11.4	10.7	95.9	2.0	11.8	9.8	8.6	93.7	1.7	14.4	12.7	10.7	96.1

fit doing only slightly better. All three fits had very accurate 95% coverage. Because the TP fit was stable and consistent in the estimate of $\beta(g,t)$ regardless of the true surface type, we again suggest using this fit in practice when the true structure of the surface is unknown.

Comparison to Local Splines

We performed a small simulation to compare our TP estimation to the penalized local least squares splines estimation in Section 2.2 of Cardot and Sarda (2008). We generated $n = 100$ or 200 Gaussian responses from

$$Y_i = \beta_0(G_i) + \int X_i(t)\beta(G_i,t)\,dt + \varepsilon_i,$$

where $\beta_0(g)$, $X_i(t)$, and ε_i are the same as in Section 4.6.1. We used only the EB surface for $\beta(g,t)$ and omit the parametric covariates. For estimation purposes, because Cardot and Sarda (2008) only consider smooth, densely observed functional covariates we fit their model and our TP model using the true $X_i(t)$ observed on the dense grid of 41 points. We evaluated the estimates for each function using the integrated squared bias, MISE, and Monte Carlo variance described in Section 4.6.1. Since Cardot and Sarda (2008) do not present methodology for finding the variances of their estimators, we did not compute the estimated variance or integrated coverage.

The results of this small study are presented in Table 4.3. For the estimates of $\beta_0(g)$, we found that the local splines fit had three times the relative bias for $n = 100$ and five times the relative bias for $n = 200$. However the overall relative bias was very small ($0.017/2.93 = 0.5\%$), thus both methods performed well in terms of integrated bias. We found that the local method had approximately nine times the variability as our TP fit for both sample sizes.

For the estimation of $\beta(g,t)$, the bias of the two fits were comparable, with the local splines estimator doing slightly better for the smaller sample size. The relative bias for both fits was about 6% for the smaller sample size, and about 3% for the larger sample size. There was about half the amount of empirical variability in the TP estimate than in the local splines fit. Both methods performed reasonably well in terms of estimating the residual variance.

4.6.3 Testing

We conducted another small simulation study to assess the Type I error and power of the naive test for interaction proposed in Section 4.4. We chose 4 settings in which the responses were either Gaussian or Binomial, and the true $\beta(g,t)$ surface had either a linear form or an EB

Table 4.3: Integrated squared bias (SB), mean integrated squared error (MSE), and integrated MC empirical variance (Vem) for $b_0(g)$ and $\beta(g,t)$ when the functional covariate is densely observed. We fit the general model to the EB surface using Cardot and Sarda (2008)'s estimation method (Car) and our proposed tensor product estimation method (TP), and estimated the residual error. All values in the table are multiplied by 100.

		$\beta_0(g)$			$\beta(g,t)$			Residual
		SB	MSE	Vem	SB	MSE	Vem	
$n = 100$	Car	1.7	42.7	41.1	4.2	26.0	21.8	116.0
	TP	0.5	5.3	4.8	4.5	14.5	10.1	89.5
$n = 200$	Car	1.1	22.5	21.4	2.3	13.0	10.7	106.5
	TP	0.2	2.3	2.1	1.9	7.9	6.0	93.3

form. We chose $n = 200$ and $n = 500$ for the Gaussian response and the Binomial responses, respectively. For the linear surface, we let $\beta(g,t) = t + 2g\sin(\pi t)$ and for the EB surface, we used the same one as in Section 4.6.1 which can be written as $\beta(g,t) = \sin(\pi t) + g\sin(\pi t) + g^2\cos(\pi t)$. Thus both surfaces are decomposed into the purely main effects of X and the purely interaction effects of X and G. All other settings for γ, β_0, etc. were the same as in Section 4.6.1 and we only used the dense grid of t.

To test the null hypothesis that there is no interaction effect, we fit the model

$$g(\mu_i) = \mathbf{W}_i^T\boldsymbol{\gamma} + \beta_0(G_i) + \int X_i(t)\left\{\beta_{\text{M}}(t) + a\beta_{\text{I}}(G_i,t)\right\}dt,$$

where a is used to determine the size and power of the test. Thus when $a = 0$, the null is true and we can determine the size of the test, and for increasing values of a, we can assess the power. We estimated the parameters of the model using the linear fit and TP fit and arbitrarily chose either $J = K = L = 5$ basis functions or $J = K = L = 9$ basis functions for the two fits.

The results of our testing simulations are presented in Figure 4.1. The solid lines correspond to the case where 5 basis functions were used and the dashed lines correspond to 9 basis functions. The red lines are the linear fit and the green indicate the TP fit. We found that none of the fits had the proper Type I error and that the Wald test was quite conservative for the TP fit. The linear fit had uniformly more power than the TP fit, but this is due to the number of parameters being tested in each fit. It is clear in the plots that the power of the test is affected by the number of basis functions chosen, especially for the TP case. This is because when $K = L = 5$, then $\dim(\boldsymbol{\eta}_{\text{I}}) = 5$ for the linear fit, compared to $\dim(\boldsymbol{\eta}_{\text{I}}) = 25$ for the TP fit. When $K = L = 9$, then $\dim(\boldsymbol{\eta}_{\text{I}}) = 9$ and $\dim(\boldsymbol{\eta}_{\text{I}}) = 81$ for the linear and TP fit respectively. The degrees of freedom increase much more rapidly for the TP fit as the number of basis functions increase, compared to the linear fit. When the responses were binary, we saw

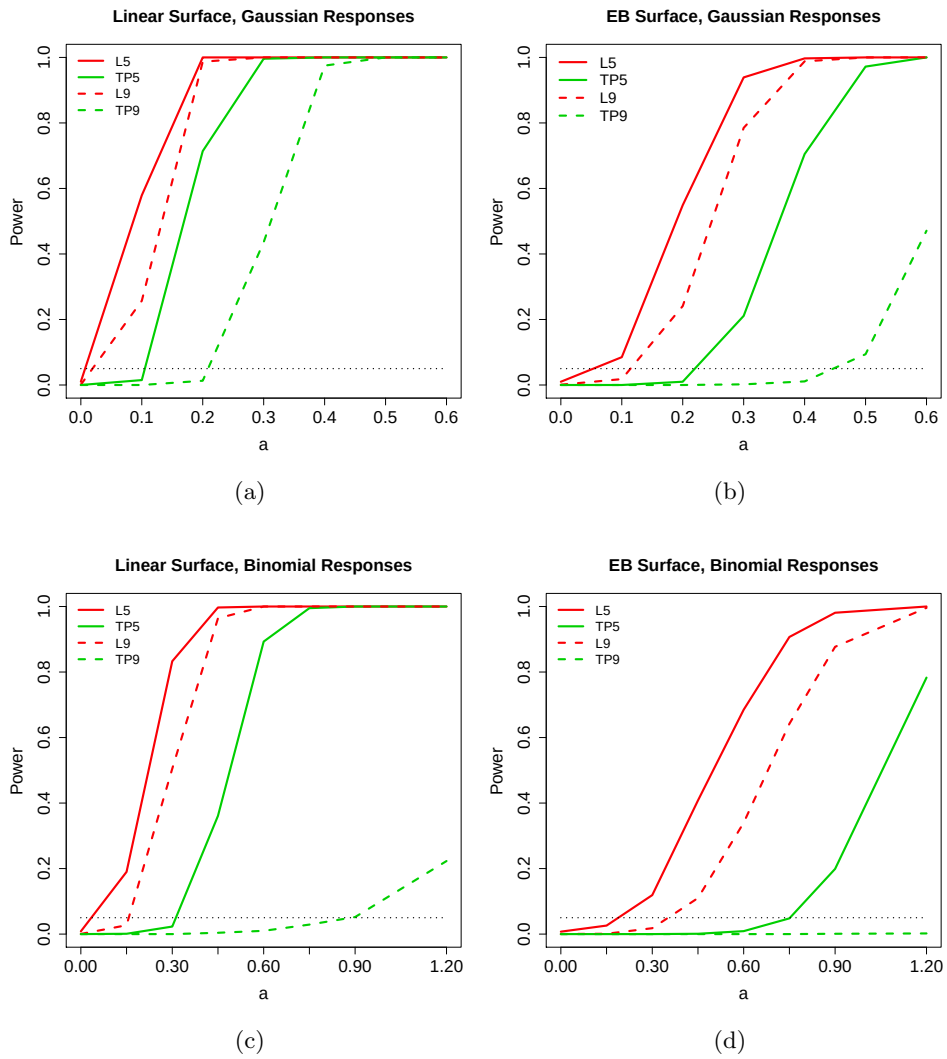

Figure 4.1: Power calculations for the linear fit (red) and TP fit (green) for $K = L = 5$ basis functions (solid) and $K = L = 9$ basis functions (dashed). The top panels show the results for the Gaussian responses, the bottom panels are the binary responses. In the left panels, the true surface was linear and in the right panels, the true surface had an EB structure.

a decrease in power for both fits and had to use a larger scale of a. We still found the linear fit to be uniformly more powerful than the TP fit, and the TP fit had very little power when the $K = L = 9$ basis functions were used for the EB surface.

We stress that the power issues in this section are due to the testing procedure and not the proposed estimation methods we have presented. In the context of this chapter, the Wald test can be unreliable and any results based on it should be viewed with some degree of skepticism. This small simulation provides evidence that testing in this framework requires much further examination.

4.7 Data Analysis

To demonstrate the methods in this chapter, we apply it to Tecator data used by several others (Ferraty and Vieu, 2006; Aneiros-Pérez and Vieu, 2006; Maity and Huang, 2012) available at http://lib.stat.cmu.edu/datasets/tecator. The data consist of percentages of moisture (water), fat, and protein content determined by analytical chemistry in finely chopped meat samples. In addition, for each meat sample, a 100-channel spectrum of absorbances were measured using a Tecator Infratec Food and Feed Analyzer. For further details, see Ferraty and Vieu (2006). The goal is to study the effects of absorbances and moisture content on fat content while adjusting for protein and the interaction between absorbances and moisture. Therefore, we denote the fat content as Y_i, the protein content as W_i, the moisture content as G_i, and the absorbances as $X_i(t_m)$ for $m = 1, ..., 100$. We are interested in modeling these data, estimating the parameters of the models, and testing for the interaction effect.

As suggested in the data description at the website above, we applied our methods to a subset of the data and used the remaining data for prediction purposes. Thus, of the 215 samples, we used 172 to estimate the parameters of the two models and for hypothesis testing purposes. We first fit the linear VCM in (4.1) and estimated the parameters using the methods described in Section 4.2, then fit the general model in (4.4) and used the EB and TP estimation methods in Section 4.3. The results are presented in Figure 4.2. The estimates of $\beta_0(g)$ were very similar for all three estimation procedures with the linear and TP fits being virtually identical. The shape of the estimates of $\beta(g, t)$ were also similar for the linear and TP fits and suggests that the true surface is linear in both G and t. The EB estimate of the surface was flatter than the other fits and had more curvature, and points to a linear relationship in G and a slightly nonlinear relationship in t. The color in the bivariate plots indicates the areas of the surface that were significantly different from zero, based on the corresponding 95% confidence surfaces. Blue areas indicate that the surface was positive and significantly different from zero, green areas mean positive but not significant, yellow areas are negative but not significant, and

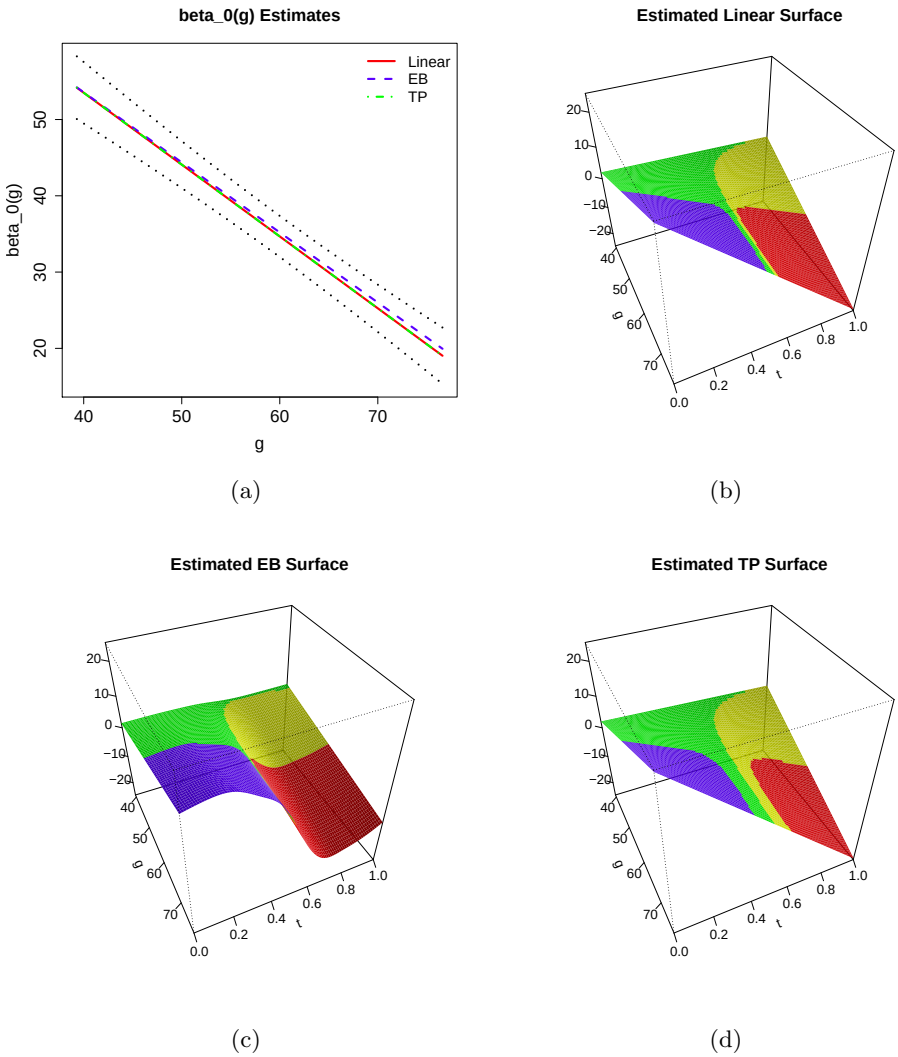

Figure 4.2: Estimated $\beta_0(g)$ (a) and estimated $\beta(g,t)$ surface from the linear fit (b), EB fit (c), and TP fit (c) for the Tecator data. The widest pointwise confidence bands were used in (a) corresponding to the EB fit. For the surface plots, blue indicates the surface values are positive and significantly different than 0, green is positive but not significant, yellow is negative but not significant, and red is negative and significant.

red areas are negative and significantly different from zero. Lastly, we found $\hat{\gamma} = -0.751$ and $\text{SE}(\hat{\gamma}) = 0.079$ to be the same up to the sixth decimal place for all three estimation schemes.

Next, we fit the linear and general models with protein and moisture reversed. That is, we fit $y_i = G_i\gamma + \beta_0(W_i) + \int X_i(t)\beta(W_i, t)\, dt + \varepsilon_i$ and these results are given in Figure 4.3. In this model, the estimates of $\beta_0(w)$ were different across the three fits but all three estimates fell within the combined confidence bands. The estimates for the bivariate surface were also different for each estimation scheme and we found larger significant regions of the surfaces than we did in the previous model. Again we see that the true surface is likely linear in both W and t for the linear and TP fits, and seems to be linear in W and nonlinear in t for the EB fit. The estimate and standard error of $\hat{\gamma}$ was -1.087 and 0.028, respectively for the linear fit. For the EB and TP fits, we found that $\hat{\gamma} = -1.092$ and $\text{SE}(\hat{\gamma}) = 0.027$ were the same up to fourth decimal place.

We then used the Wald statistic to test for effects in the linear model and in the general model using TP estimation. In both models, the interaction effect of moisture and absorbances was not found to be significant ($p = 0.352$ and 0.957 for the linear and TP fit, respectively). Given than the interaction was not significant, we then tested for the main effects of X and G by fitting the model in (4.10). The main effect of G was found to be highly significant ($p << 0.001$), whereas the p-value for testing the effect of X was 0.087. Because p-values close to the decision threshold need to be considered with care when using this naive test, it is difficult to make any inference on this result.

Next we applied our testing procedure to the model where protein and moisture were switched. For the linear model, the test for significant interaction resulted in a p-value of 0.041. Again, we can not reliably draw any conclusions based on this value. For the TP case, because the test for interaction was not significant ($p = 0.317$) we fit the main effects model an tested for the effects of G and X. The effect of G was highly significant ($p << 0.001$) whereas the test for X was difficult to call ($p = 0.016$).

Finally, to evaluate our models of interest, we compared them to other models using the prediction MSE criterion defined as PMSE= $\sum_{i=1}^{43}(\hat{Y}_i - Y_i)^2/43$ and these results are presented in Table 4.4. The lowest PMSEs occurred using the switched general model where moisture content was modeled parametrically and protein content was modeled nonparametrically. These values were 1.71 and 1.78 for the TP and EB fits respectively. When using the original models, the linear fit and the TP fit had the same PMSE (1.78) whereas the EB fit had a slightly higher PMSE (1.85). In general, we found that modeling the main effect of moisture nonparametrically yielded lower PMSE scores than a parametric fit, and that any model that did not contain moisture content as a covariate did not predict the responses well.

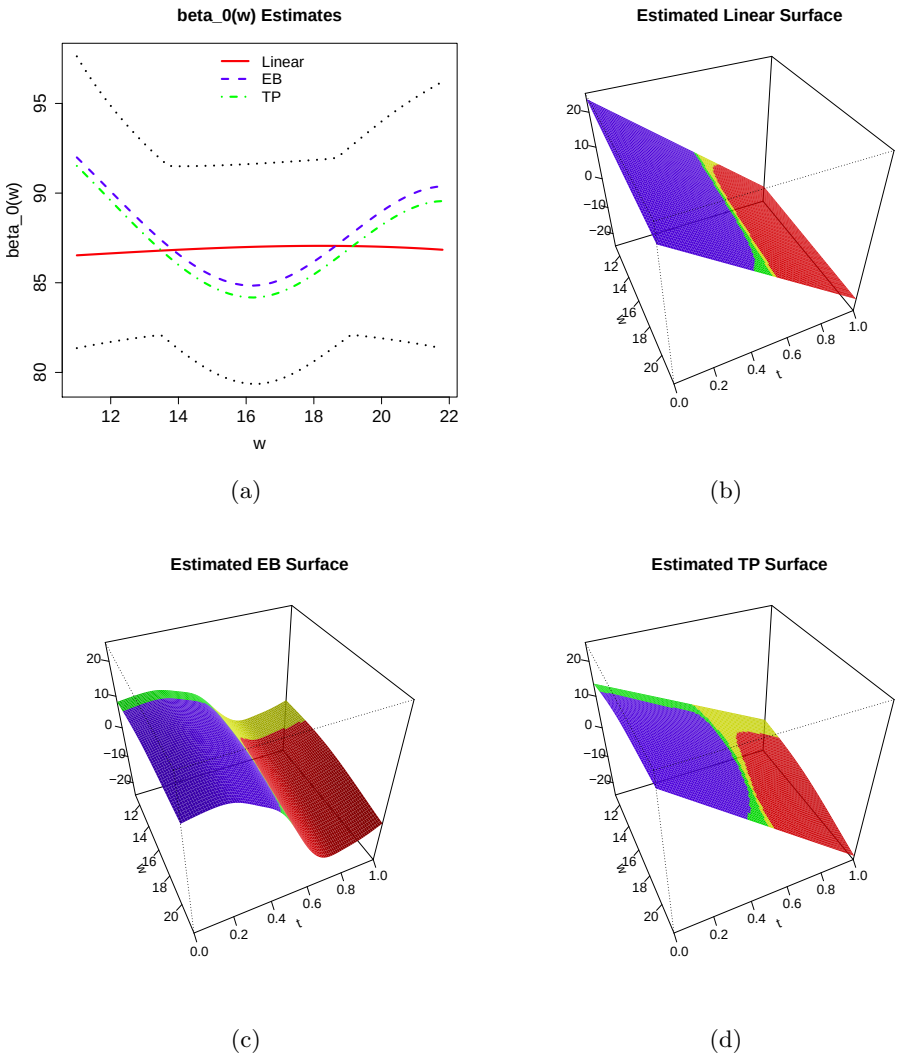

Figure 4.3: Estimated $\beta_0(w)$ (a) and estimated $\beta(w,t)$ surface from the linear fit (b), EB fit (c), and TP fit (c) for the Tecator data. The widest pointwise value was used for the confidence bands in (a). For the surface plots, blue indicates the surface values are positive and significantly different than 0, green is positive but not significant, yellow is negative but not significant, and red is negative and significant.

Table 4.4: Prediction MSE for the Tecator data for several different models as a comparison to our models of interest (the first four). For the general models the two PMSE values correspond to EB estimation/TP estimation. Recall that G_i is moisture content and W_i is protein content.

Model: $E(Y_i \mid W_i, G_i, X_i) =$	PMSE
$W_i\gamma + \beta_0(G_i) + \int X_i(t)\beta(G_i, t)\, dt$	1.85/1.78
$W_i\gamma + \beta_0(G_i) + \int X_i(t)\beta_1(t)\, dt + G_i \int X_i(t)\beta_2(t)\, dt$	1.78
$G_i\gamma + \beta_0(W_i) + \int X_i(t)\beta(W_i, t)\, dt$	1.77/1.71
$G_i\gamma + \beta_0(W_i) + \int X_i(t)\beta_1(t)\, dt + W_i \int X_i(t)\beta_2(t)\, dt$	1.80
$W_i\gamma + \beta_0(G_i)$	2.05
$G_i\gamma + \beta_0(W_i)$	2.26
$\alpha + W_i\gamma + G_i\beta$	2.26
$\alpha + W_i\gamma + G_i\beta + W_iG_i\delta$	2.22
$\alpha + W_i\gamma + \int X_i(t)\beta(t)\, dt$	36.39
$\alpha + G_i\gamma + \int X_i(t)\beta(t)\, dt$	4.35

4.8 Discussion

In this chapter we established several methodologies for modeling the joint effect of a scalar covariate and a functional covariate and adjusting to for interaction between them. We proposed a varying coefficient model in the functional framework for when the scalar covariate is categorical, and presented a more general model using a bivariate surface function for continuous covariates. Two estimation schemes were presented, each with their own benefits and drawbacks. We conducted a simulation study to compare the methods in a variety of settings, and found that one method was consistent and stable and thus is probably the best choice in real world applications.

The testing procedure presented in this chapter is naive at best. Testing in this framework is difficult, and we encountered several issue using the Wald statistic. First, the Wald test does not account for estimation of the smoothing parameter, so any resulting p-values close to the decision threshold will be unreliable. Because of this, we had difficulty drawing any conclusions when the test was applied to a real data example. Secondly, we found the test to be somewhat conservative in simulations, and it did not have the proper Type I error. Lastly, there may be some identifiability issues with the bivariate surface for testing in general, however more research is needed in order to determine for certain if this phenomenon occurs in practice. Thus we again stress that developing accurate and reliable testing procedures is an open problem that requires further investigation.

Finally, all of the estimation methods presented in this chapter were spline based, but the estimation schemes given in the two previous chapters can be extended to this functional framework. Recall that the methods in Sections 4.2 and 4.3 require estimation of unknown

parameter functions of a single argument (e.g. β_0, β_1, b_k, etc). This was exactly the problem in the previous chapters where kernel machine theory was used for estimation in Chapter 2 and local polynomial fitting with a parametric guide was used in Chapter 3. These estimation tools can easily replace the spline based procedures that were used, and can possibly extended to the case where the unknown function has multiple arguments.

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis*. 2nd edition, New York: Wiley.

Aneiros-Pérez, G. and Vieu, P. (2006), Semi-functional partial linear regression, *Statistics and Probability Letters* **76**, 1102–1110.

Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L. E., and Morris, J. S. (2010), Bayesian random segmentation models to identify shared copy number aberrations for array CGH data, *Journal of the American Statistical Association* **105**, 1358–1375.

Bauer, C. R., Shankaran, S., Bada, H. S., Lester, B., Wright, L. L., Krause-Steinrauf, H., Smeriglio, V. L., Finnegan, L. P., Maza, P. L., and Verter, J. (2002), The maternal lifestyle study: drug exposure during pregnancy and short-term maternal outcomes, *American Journal of Obstetrics and Gynecology* **186**, 487–495.

Brumback, B. and Rice, J. A. (1998), Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion), *Journal of the American Statistical Association* **93**, 961–994.

Buhmann, M. D. (2003), *Radial Basis Functions: Theory and Implementations*, Cambridge: Cambridge University Press.

Burges, C. J. (1998), A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery* **2**, 121–167.

Cai, Z., Fan, J., and Li, R. (2000), Efficient estimation and inferences for varying-coefficient models, *Journal of the American Statistical Association* **95**, 888–902.

Cao, Y., Lin, H., Wu, T. Z., and Yu, Y. (2010), Penalized spline estimation for functional coefficient regression models, *Computational Statistics and Data Analysis* **54**, 891–905.

Cardot, H. (2007), Conditional functional principal components analysis, *Scandinavian Journal of Statistics* **34**, 317-335.

Cardot, H. and Sarda, P. (2008), Varying-coefficient functional linear regression models, *Communications in Statistics - Theory and Methods* **37**, 3186–3203.

Carey, J. R., Liedo, P., Müller, H.-G., Wang, J.-L., and Chiou, J.-M. (1998), Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of mediterranean fruit fly females, *Journal of Gerontology Series A: Biological Sciences* **53**, B245–B251.

Chen, R. and Tsay, R. S. (1993), Functional-coefficient autoregressive models, *Journal of the American Statistical Association* **88**, 298–308.

Cleveland, W., Grosse, E., and Shyu, W. (1991), Local regression models, *Statistical Models in S* (eds J. Chambers and T. Hastie), Pacific Grove, CA: Wadsworth and Brooks/Cole, 309–376.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009), Generalized multilevel functional regression, *Journal of the American Statistical Association* **104**, 1550–1561.

Crambes, C., Kneip, A., and Sarda, P. (2009), Smoothing splines estimators for functional linear regression, *The Annals of Statistics* **37**, 35–72.

Cristianini, N. and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge, U.K.: Cambridge University Press.

Das, A., Poole, W. K., and Bada, H. S. (2004). A repeated measures approach for simultaneous modeling of multiple neurobehavioral outcomes in newborns exposed to cocaine in utero, *American Journal of Epidemiology* **159**, 891–899.

Davenport, C. A., Maity, A., and Wu, Y. (2013), Parametrically guided estimation in nonparametric varying coefficient models with quasi-likelihood, In Progress.

Di, C., Crainiceanu, C., Caffo, B., and Punjabi, N. (2009), Multilevel functional principal component analysis, *Annals of Applied Statistics* **3**, 458–488.

Eilers, P.H.C. and Marx, B.D. (1996), Flexible smoothing with B-splines and penalities, *Statistical Science* **11**, 89-102.

Fan, J. (1993), Local linear regression smoothers and their minimax efficiencies, *The Annals of Statistics* **21**, 196–216.

Fan, J., Farmen, M., and Gijbels, I. (1998), Local maximum likelihood estimation and inference, *Journal of the Royal Statistical Society Series B* **60**, 591–608.

Fan, J., Maity, A., Wang, Y., and Wu, Y. (2013), Parametrically guided generalised additive models with application to mergers and acquisitions data, *Journal of Nonparametric Statistics* **25**, 109–128.

Fan, J., Wu, Y., and Feng, Y. (2009), Local quasi-likelihood with a parametric guide, *The Annals of Statistics* **37**, 4153–4183.

Fan, J. and Zhang, W. (1999), Statistical estimation in varying coefficient models, *The Annals of Statistics* **27**, 1491–1518.

Fan, J. and Zhang, W. (2008), Statistical methods with varying coefficient models, *Statistics and Its Interface* **1**, 179–195.

Ferraty, F., González-Manteiga, W., Martínez-Calvo, A., and Vieu, P. (2012), Presmoothing in functional linear regression, *Statistica Sinica* **22**, 69–94.

Ferraty, F. and Vieu, P., (2006), *Nonparametric Functional Data Analysis*, New York: Springer.

Gabrys, R., Horváth, L., and Kokoszka, P. (2010), Tests for error correlation in the functional linear Model, *Journal of the American Statistical Association* **105**, 1113–1125.

Glad, I. K. (1998), Parametrically guided non-parametric regression, *Scandinavian Journal of Statistics* **25**, 649–668.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013), Corrected confidence bands for functional data using principal components, *Biometrics* **69**, 41–51.

Gu, C. and Kim, Y.-J. (2002), Penalized likelihood regression: general formulation and efficient approximation, *The Canadian Journal of Statistics* **30**, 619–628.

Hastie, T. and Tibshirani, R. (1993), Varying-coefficient models, *Journal of the Royal Statistical Society Series B* **55**, 757–796.

Hjort, N. L. and Glad, I. K. (1995), Nonparametric density estimation with a parametric start, *The Annals of Statistics* **23**, 882–904.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008), Kernel methods in machine learning, *The Annals of Statistics* **36**, 1171–1220.

Holdeman, J. T. (1969), A method for the approximation of functions defined by formal series expansions in orthogonal polynomials, *Mathematics of Computation* **23**, 275-287.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998), Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika* **85**, 809–822.

Huang, J. Z. and Shen, H. (2004), Functional coefficient regression models for nonlinear time series: A polynomial spline approach, *Scandinavian Journal of Statistics* **31**, 515–534.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008), A powerful and flexible multilocus association test for quantitative traits, *The American Journal of Human Genetics* **82**, 386–397.

Laukaitis, A. (2008), Functional data analysis for cash flow and transactions intensity continuous-time prediction using Hilbert-valued autoregressive processes, *The European Journal of Operational Research* **185**, 1607–1614.

Lederman, M. M., Connick, E., Landay, A., Kuritzkes, D. R., Spritzler, J., St. Clair, M., Kotzin, B. L., Fox, L., Chiozzi, M. H., Leonard, J. M. et al. (2002), Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir: results of AIDS Clinical Trials Group Protocol 315, *The Journal of Infectious Diseases* **179**, 70–79.

Liang, H., Wu, H., and Carroll, R. J. (2003), The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement Error, *Biostatistics* **4**, 297–312.

Liang, K. Y., and Zeger, S. L. (1986), Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keef, R. S. E., Davis, S. M., Davis, C. E., Lebowitz, B. D., Severe, J., and Hsiao, J. K. (2005), Effectiveness of antipsychotic drugs in patients with chronic schizophrenia, *The New England Journal of Medicine* **353**, 1209–1223.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Sinha, D., Parzen, M., and Lipshultz, S. (2009), Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data, *Journal of the Royal Statistical Society Series A* **172**, 3–20.

Liu, D., Ghosh, D., and Lin, X. (2008), Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models, *BMC Bioinformatics* 9:292.

Liu, D., Lin, X., and Ghosh, D. (2007), Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models, *Biometrics* **63**, 1077–1088.

Ma, S., Yang, L., Romero, R., and Cui, Y. (2011), Varying coefficient model for geneenvironment interaction: a non-linear look, *Bioinformatics* **27**, 2119–2126.

Maity, A. and Huang, J. Z. (2012), Partially linear varying coefficient models stratified by a functional covariate, *Statistics and Probability Letters* **82**, 1807–1814.

Maity, A., Sullivan, P. F., and Tzeng, J. Y. (2012), Multivariate phenotype association analysis by marker-set kernel machine regressions, *Genetic Epidemiology* **36**, 686–695.

Martins-Filho, C., Mishra, S., and Ullah, A. (2008), A class of improved parametrically guided nonparametric regression estimators, *Econometric Reviews* **27**, 542–573.

Maslova, I., Kokoszka, P., Sojka, J., and Zhu, L. (2010), Statistical significance testing for the association of magnetometer records at high-, mid- and low latitudes during substorm days, *Planetary and Space Science* **58**, 437–445.

Meiring, W. (2007), Oscillations and time trends in stratospheric ozone levels: a functional data analysis approach *Journal of the American Statistical Association* **102**, 788–802.

Müller, H.-G. (1987), Weighted local regression and kernel methods for nonparametric curve fitting, *Journal of the American Statistical Association* **82**, 231–238.

Müller, H.-G. and Stadtmüller, U. (2005), Generalized functional linear models, *The Annals of Statistics* **33**, 774–805.

Nam, D. and Kim, S. Y. (2008), Gene-set approach for expression pattern analysis, *Briefings in Bioinformatics* **9**, 189–197.

Nelder, J. A. and Wedderburn, R. W. M (1972), Generalized linear models, *Journal of the Royal Statistical Society Series A* **135**, 370–384.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J. (1986), Automatic smoothing of regression functions in generalized linear models, *Journal of the American Statistical Association* **81**, 96–103.

Pan, K. H., Lih, C. J., and Cohen, S. N. (2005), Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays, *Proceedings of the National Academy of Science of the United States of America* **102**, 8961–8965.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, 2nd edition, New York: Springer.

Staicu, A.-M., Crainiceanu, C. M., Reich, D. S., and Ruppert, D. (2012), Modeling functional data with spatially heterogeneous shape characteristics, *Biometrics* **68**, 331–343.

Staniswalis, J. G. and Lee, J. J. (1998), Nonparametric regression analysis of longitudinal data, *Journal of the American Statistical Association* **93**, 1403–1418.

Stankiewicz, P. and Lupski, J. R. (2010), Structural variation in the human genome and its Role in disease, *Annual Review of Medicine* **61**, 437–455.

Sullivan, P. F., Lin, D., Tzeng, J. Y., van den Oord, E., Perkins, D., Stroup, T.S., Wagner, M., Lee, S., Wright, F. A., Zou, F., Liu, W., Downing, A. M., Lieberman, J., and Close, S. L. (2008), Genomewide association for schizophrenia in the CATIE study: results of Stage 1, *Molecular Psychiatry* **13**, 570–584.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J. (2002), *Least Squares Support Vector Machines*, Singapore: World Scientific.

Torres, J. M., Nieto, P. J. G., Alejando, L., and Reyes, A. N. (2011), Detection of outliers in gas emissions from urban areas using functional data analysis, *Journal of Hazardous Materials* **186**, 144–149.

Ullah, S. and Finch, C. F. (2010), Functional data modelling approach for analysing and predicting trends in incidence rates - an application to falls injury, *Osteoporosis International* **21**, 2125–2134.

Ullah, S. and Finch, C. F. (2013), Applications of functional data analysis: a systematic review, *BMC Medical Research Methodology* **13**:43.

Vapnik, V.N. (1998), *Statistical Learning Theory*, New York: Wiley.

Wedderburn, R. W. M. (1974), Quasi-likelihood functions, generalized linear models, and the GaussNewton method, *Biometrika* **61**, 439–447.

Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*, Boca Raton, FL: Chapman and Hall/CRC.

Wood, S. N. (2011), Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, *Journal of the Royal Statistical Society Series B* **73**, 3–36.

Wu, Y., Fan, J., and Müller, H.-G. (2010), Varying-coefficient functional linear regression, *Bernoulli* **16**, 730-758.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011), Rare variant association testing for sequencing data using the sequence kernel association test (SKAT), *The American Journal of Human Genetics* **89**, 82–93.

Wu, P. S. and Müller, H.-G. (2010), Functional embedding for the classification of gene expression profiles, *Bioinformatics* **26**, 509–517.

Wu, H. and Wu, L. (2002), Identification of significant host factors for HIV dynamics modelled by non-linear mixed-effects models, *Statistics in Medicine* **21**, 753–771.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association* **100**, 577–590.

Yolken, R. H., Torrey, E. F., Lieberman, J. A., Yang, S., and Dickerson, F. B. (2011), Serological evidence of exposure to herpes simplex virus Type 1 is associated with cognitive deficits in the CATIE schizophrenia sample, *Schizophrenia Research* **128**, 61–65.